REPORT

FOR

FINAL PROJECT: DATA SCIENCE SALARY ANALYSIS FROM YEAR 2020 TO 2023

Prepared by:

SYED AZHARI BIN SYED RIDHUAN

29th JUNE 2023

TABLE OF CONTENTS

No.	Content	Page Number
1.0	**Report Overview**	**3**
2.0	**Suitable Dataset For the Analysis**	**4**
3.0	**Exploratory Data Analysis**	**6**
3.1	Dataset Cleaning	6
3.2.1	Univariate Pie Chart	9
3.2.2	Univariate Bar Chart	12
3.2.3	Bivariate Line Chart	16
3.2.4	Bivariate Grouped Bar Chart	20

1.0 Report Overview

This data science-related study aims to analyze and provide insights into the salaries within the field of data science for the year 2020 to 2023. By leveraging the dataset, we will examine various variables such as work year, experience level, employment type, job title, salary, salary currency, salary in USD, employee residence, remote ratio, company location, and company size. The focus of this study will align with the education category by exploring the relationship between educational factors and salaries in the data science field.

1.1 Study Purpose

The purpose of this study is to gather information and analyze the salaries within the field of data science specifically for the year 2020 to 2023. By examining the dataset variables, the study aims to understand the salary landscape within the data science industry and its connection to educational factors. This study aims to provide valuable insights into the salaries offered in the data science field, focusing on the educational aspects that contribute to salary variations.

1.2 Objectives

1.2.1 Analyze Salary Trends

The objective is to identify and analyze trends in data science salaries for the year 2020 to 2023. By examining the work year variable, we can identify any notable changes or patterns in salary levels over time, providing insights into the evolving nature of data science salaries.

1.2.2 Explore Factors Influencing Salaries

This objective aims to identify and examine the factors that influence data science salaries. Variables such as experience level, employment type, job title, and company size in the dataset can help us to analyse and understand how these factors relate to salary variations. Analyzing these relationships can provide insights into the impact of educational factors on salaries within the data science field.

1.2.3 Understand Salary Distributions

Another objective is to gain insights into the distribution of salaries across different industries or job roles within the data science field. By examining the variables of the dataset in relation to job titles and company locations, we can identify salary ranges and distributions specific to different sectors or roles. This analysis can contribute to understanding the salary landscape within the education-focused data science industry.

1.3 Target Audience

1.3.1 Aspiring Data Scientists

This group of fresh grads or employees can benefit from the study by gaining insights into salary trends and understanding the factors that can influence data science salaries. From this knowledge, it can help them make informed decisions regarding their education and career paths.

1.3.2 Data Science Educators

Educators can use the study findings to understand salary distributions and trends in the data science industry. They can provide feedback and inform the curriculum development and program planning in data science related programmes to align with industry demands and prepare students for competitive salaries and skillsets to succeed in the field.

1.3.3 Researchers and Academics

Researchers focusing on the intersection of education and data science salaries can leverage this study for academic research and analysis. The dataset can be analysed to provide valuable information for studies related to the educational aspects of data science salaries.

1.3.4 Hiring Managers and Recruiters

Professionals or senior employees involved in hiring data science talent can benefit from this study by understanding salary benchmarks and trends within the industry. It can assist in salary negotiations and ensuring competitive compensation packages under their company or organization.

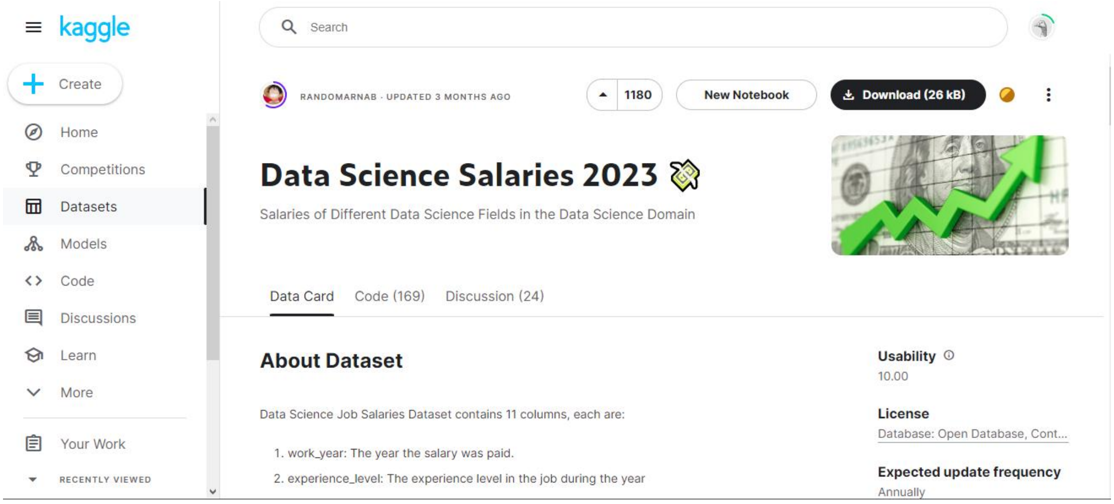
1.3.5 Policymakers

Policymakers that are interested in promoting education in the data science field and supporting the workforce development can gain insights into salary distributions and trends. This can help them to improve certain policies and initiatives aimed at fostering the growth of the data science industry.

2.0 Suitable Dataset For the Analysis

In order to conduct the analysis, Im using “Data Science Salaries 2023” dataset by RANDOMARNAB from Kaggle website. There are variables that can provide key information for analyzing and understanding data science salaries in the dataset. They cover various aspects such as temporal information, experience level, job titles, salary amounts, currencies, geographic factors, employment types, and company characteristics. By examining these variables, we can gain insights into salary trends, distributions, and factors that influence data science salaries from 2020 to 2023.

The Website URL for the dataset :
https://www.kaggle.com/datasets/arnabchaki/data-science-salaries-2023



Data Science Salaries 2023

Salaries of Different Data Science Fields in the Data Science Domain

About Dataset

Data Science Job Salaries Dataset contains 11 columns, each are:

1. work_year: The year the salary was paid.
2. experience_level: The experience level in the job during the year

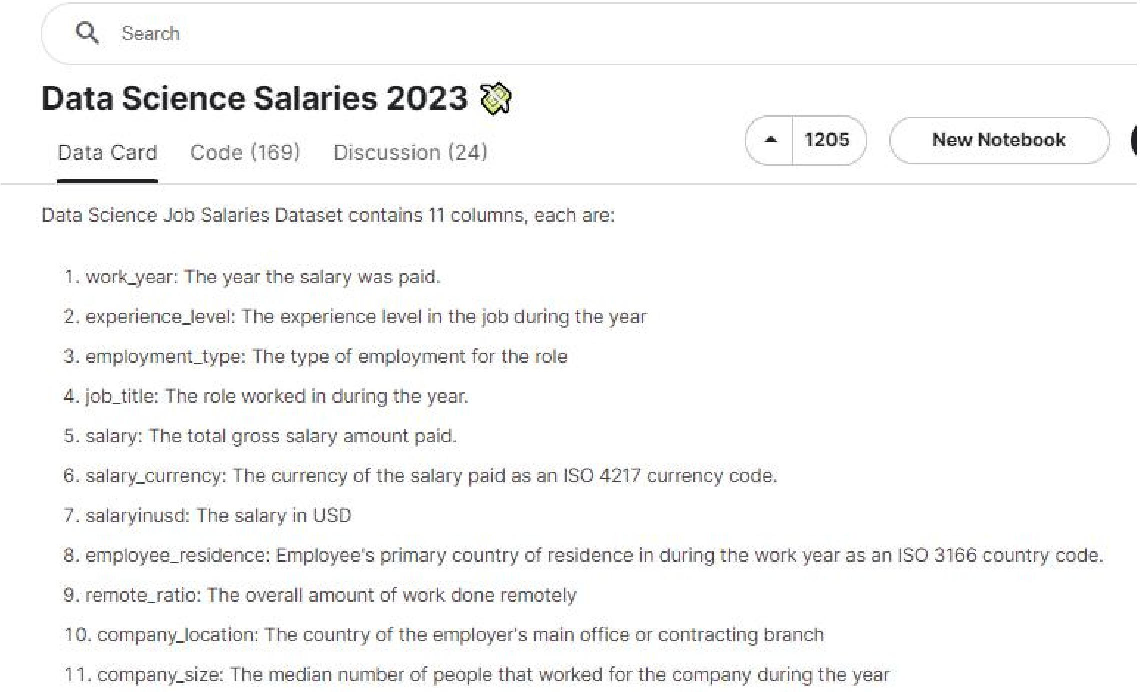
Usability
10.00

License
Database: Open Database, Cont...

Expected update frequency
Annually

Diagram 2.0.1: *The website of kaggle with the information about the dataset*

2.1 The Variables of the Dataset



Data Science Salaries 2023

Data Card Code (169) Discussion (24)

Data Science Job Salaries Dataset contains 11 columns, each are:

1. work_year: The year the salary was paid.
2. experience_level: The experience level in the job during the year
3. employment_type: The type of employment for the role
4. job_title: The role worked in during the year.
5. salary: The total gross salary amount paid.
6. salary_currency: The currency of the salary paid as an ISO 4217 currency code.
7. salaryinusd: The salary in USD
8. employee_residence: Employee's primary country of residence in during the work year as an ISO 3166 country code.
9. remote_ratio: The overall amount of work done remotely
10. company_location: The country of the employer's main office or contracting branch
11. company_size: The median number of people that worked for the company during the year

Diagram 2.1.1: *The dataset variables from kaggle website*

2.1.1 work_year

This column represents the year in which the salary was paid. It provides a temporal aspect to the dataset, allowing analysis of salary trends and changes over time.

2.1.2 experience_level

This column indicates the experience level of individuals in their job during the specified year. It could include categories such as junior, mid-level, senior, or other relevant distinctions. This variable helps understand the relationship between experience and salary.

2.1.3 employment_type

This column describes the type of employment for the role, such as full-time, part-time, contract, or freelance. It provides insights into the different employment arrangements and how they may relate to salary levels.

2.1.4 job_title

This column specifies the role that individuals worked in during the specified year. It could include job titles like data scientist, data analyst, machine learning engineer, or other relevant positions. This variable allows for analyzing salary differences based on different job roles.

2.1.5 salary

The salary column represents the total gross salary amount paid to individuals. It is a numerical variable that provides the primary focus of the dataset, allowing analysis of salary distributions, averages, and other statistical measures.

2.1.6 salary_currency

This column specifies the currency in which the salary was paid, using ISO 4217 currency codes. It helps understand the currency context for salary amounts, which can be important for comparing salaries across countries or converting them to a common currency for analysis.

2.1.7 salaryinusd

This column provides the salary amount in USD (United States Dollars). It allows for standardizing salary amounts and comparing them on a common currency basis, facilitating cross-country or international salary analyses.

2.1.8 employee_residence

This column represents the employee's primary country of residence during the work year, using ISO 3166 country codes. It provides information about the geographic distribution of data science professionals and enables analysis of salary differences based on the employee's location.

2.1.9 remote_ratio

This column indicates the overall amount of work done remotely by individuals. It represents the proportion or ratio of remote work to total work hours. This variable allows for understanding the prevalence and impact of remote work on salaries.

2.1.10 company_location

This column specifies the country of the employer's main office or contracting branch. It provides insights into the geographic distribution of companies hiring data science professionals and allows for analyzing salary differences based on the employer's location.

2.1.11 company_size

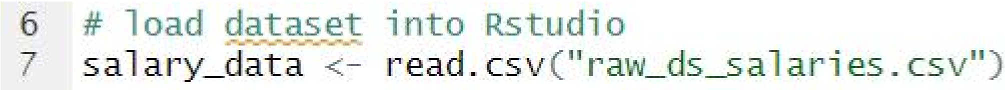
This column represents the median number of people that worked for the company during the year. It provides information about the size of the employing organizations and can help analyze salary differences based on company size.

3.0 Task 3: Conduct exploratory data analysis (EDA).

3.1 Data Cleaning

To conduct data cleaning from the “Data Science Salary in 2023” dataset, these steps were taken;

Step 1: Load the Dataset

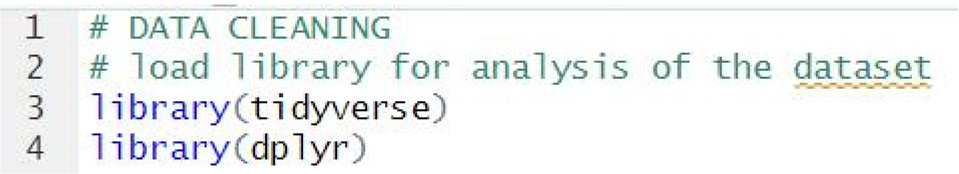


```
# load dataset into Rstudio
salary_data <- read.csv("raw_ds_salaries.csv")
```

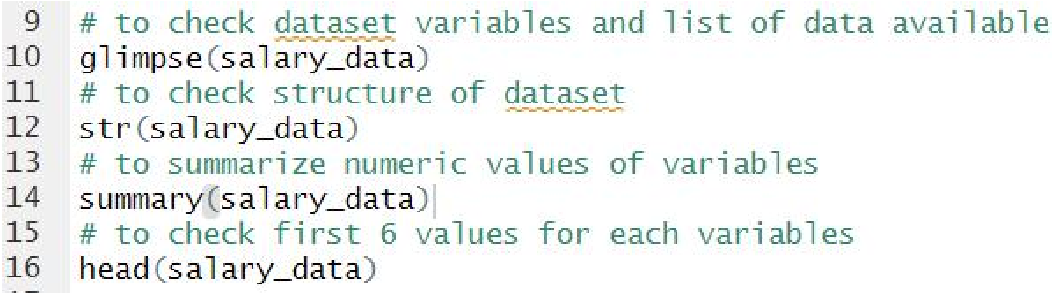
Diagram 3.1.1: *Code Snippet to load the dataset*

In this step, you will load the dataset into RStudio using the appropriate function, such as read.csv() or read_excel(). The dataset is assigned to a variable for the next step of the data cleaning process. For example, i’ve using raw_ds_salary for uncleaned version of the dataset.

Step 2: Explore the Dataset



```
# DATA CLEANING
# load library for analysis of the dataset
library(tidyverse)
library(dplyr)
```

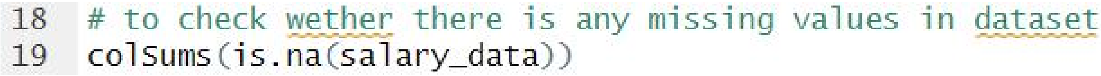


```
# to check dataset variables and list of data available
glimpse(salary_data)
# to check structure of dataset
str(salary_data)
# to summarize numeric values of variables
summary(salary_data)
# to check first 6 values for each variables
head(salary_data)
```

Diagram 3.1.2: *Code Snippet to explore the dataset*

You can explore the dataset to get a better understanding of its structure and contents. Using library such as tidyverse with built-in functions like str(), summary(), and head() or dplyr with glimpse(), you can view the structure, summary, statistics, and the first few rows of the dataset respectively. This helps you to identify the variables and potential issues such as missing or incorrect values or zero quantities in the dataset.

Step 3: Identify Missing Values

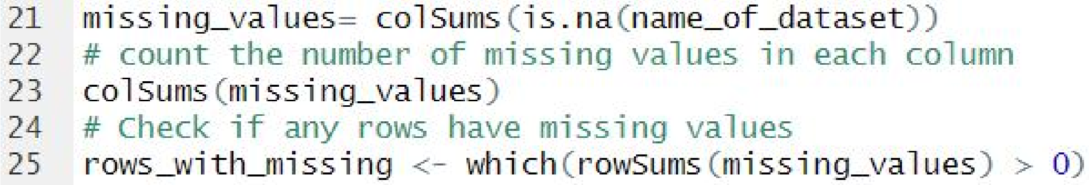


```
# to check wether there is any missing values in dataset
colSums(is.na(salary_data))
```

Diagram 3.1.3: *Code Snippet to identify missing values in the dataset*

You can use functions like is.na() or complete.cases() to identify missing values in the dataset. The is.na() function returns a logical matrix indicating which values are missing (TRUE) and which are not (FALSE). The colSums() function is then used to count the number of missing values in each column. You can also check if any rows have missing values using rowSums() and logical indexing.

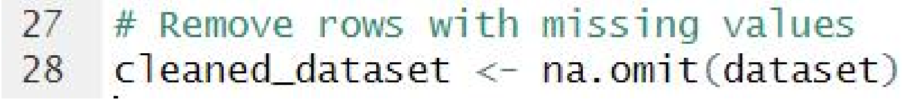
Step 4: Handle Missing Values



```
missing_values= colSums(is.na(name_of_dataset))
# count the number of missing values in each column
colSums(missing_values)
# Check if any rows have missing values
rows_with_missing <- which(rowSums(missing_values) > 0)
```

Diagram 3.1.4: *Code Snippet to check missing values in the dataset*

Remove missing values

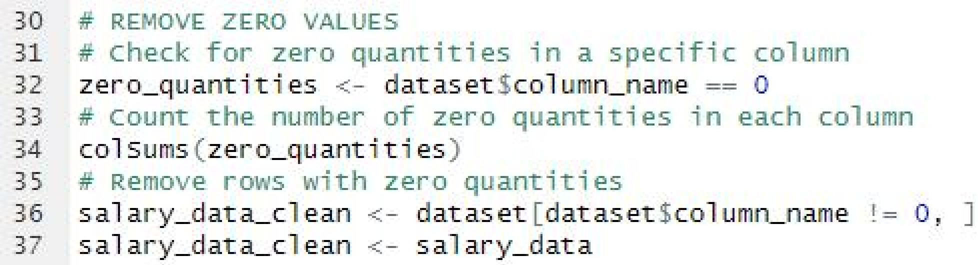


```
# Remove rows with missing values
cleaned_dataset <- na.omit(dataset)
```

Diagram 3.1.5: *Code Snippet to handle missing values in the dataset*

If the missing values are few and randomly distributed, you can remove the corresponding rows using the na.omit() function. The na.omit() function removes rows with any missing values, resulting in a cleaned dataset.

Step 5: Identify Zero Quantities

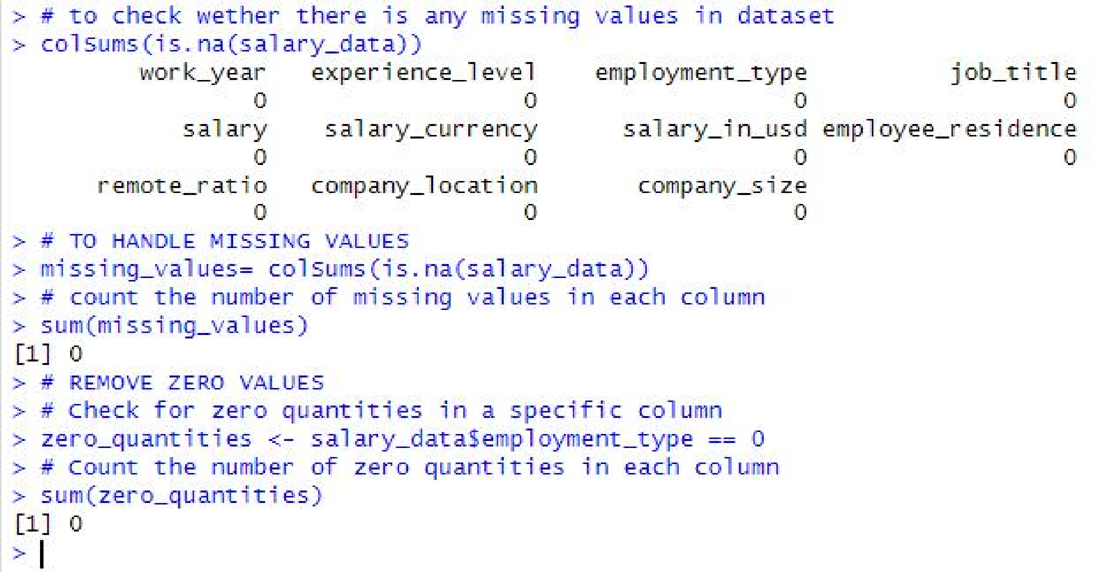


```
# REMOVE ZERO VALUES
# Check for zero quantities in a specific column
zero_quantities <- dataset$column_name == 0
# Count the number of zero quantities in each column
colSums(zero_quantities)
# Remove rows with zero quantities
salary_data_clean <- dataset[dataset$column_name != 0, ]
salary_data_clean <- salary_data
```

Diagram 3.1.6: *Code Snippet to identify quantities of zero in the dataset*

You check the relevant variables in the dataset to identify quantities of zero. The comparison dataset$column_name == 0 creates a logical vector where TRUE indicates the presence of a zero quantity.

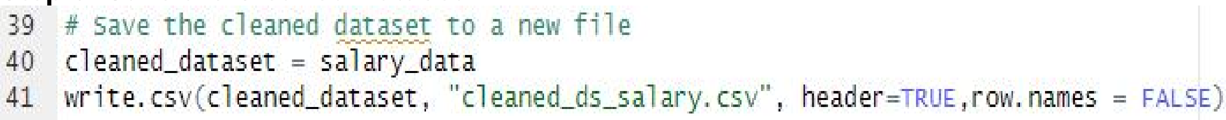
If zero quantities are inappropriate for analysis or considered outliers, you can remove the corresponding rows using logical indexing. The code dataset$column_name != 0 creates a logical vector where TRUE indicates the absence of zero quantities. For the dataset that im going to analyse, I found no missing, zero, outliers or incorrect values for each columns from previous step of analysis, so this step is not required in and I proceeded with the next step, save the cleaned dataset.



```
> # to check wether there is any missing values in dataset
> colSums(is.na(salary_data))
         work_year    experience_level     employment_type           job_title
                 0                   0                   0                   0
            salary     salary_currency       salary_in_usd employee_residence
                 0                   0                   0                   0
      remote_ratio    company_location        company_size
                 0                   0                   0
> # TO HANDLE MISSING VALUES
> missing_values= colSums(is.na(salary_data))
> # count the number of missing values in each column
> sum(missing_values)
[1] 0
> # REMOVE ZERO VALUES
> # Check for zero quantities in a specific column
> zero_quantities <- salary_data$employment_type == 0
> # Count the number of zero quantities in each column
> sum(zero_quantities)
[1] 0
> |
```

Diagram 3.1.7: *Console output of the dataset after it is inspected*

Step 6: Save the Cleaned Dataset



```
# Save the cleaned dataset to a new file
cleaned_dataset = salary_data
write.csv(cleaned_dataset, "cleaned_ds_salary.csv", header=TRUE,row.names = FALSE)
```

Diagram 3.1.8: *Code Snippet to save the cleaned dataset from missing or unconsistent values*

Finally, you can save the cleaned dataset to a new file using the write.csv() function. This creates a new CSV file with the cleaned dataset, ready for further analysis.

3.2 Univariate Analysis

3.2.1 Categorical Category: Distribution of Individual’s Experience Level In Data Science

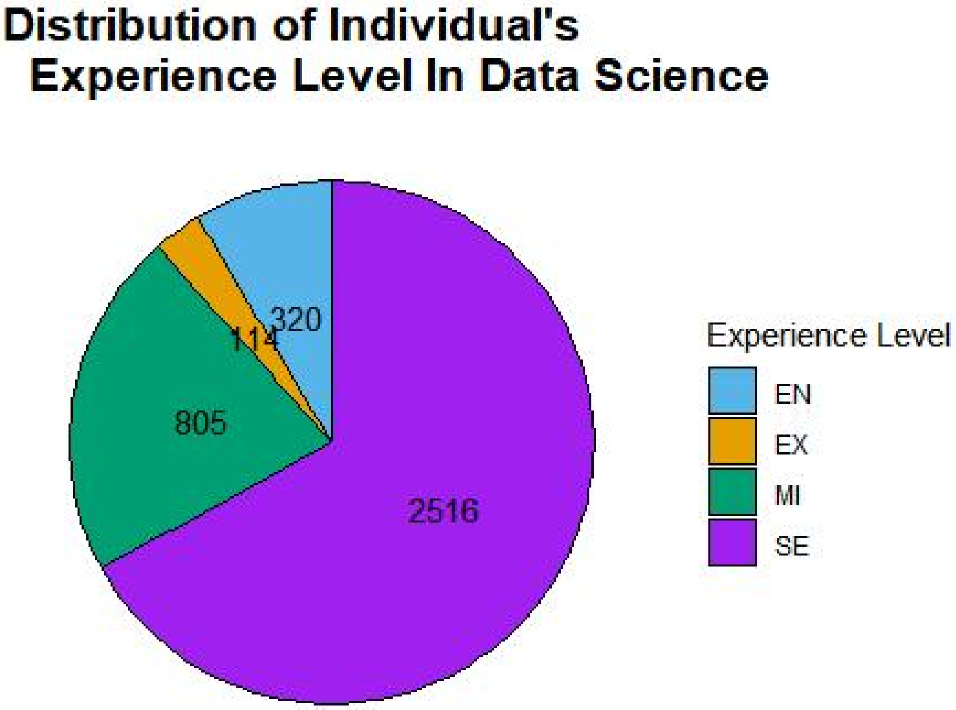
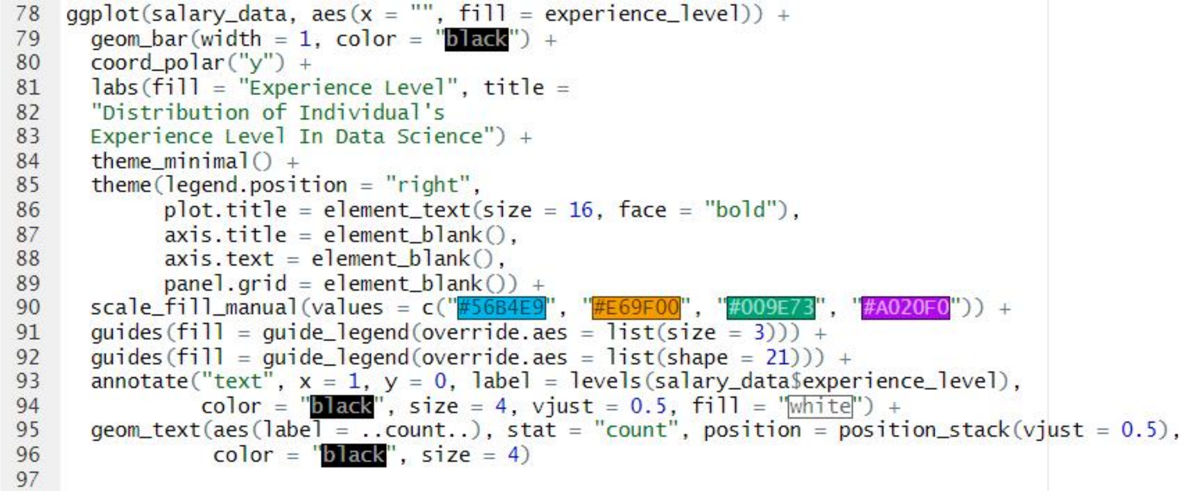


Diagram 3.2.1.0.1: *Pie Chart : Distribution of Experience Level*

Rstudio Script



```
ggplot(salary_data, aes(x = "", fill = experience_level)) +
  geom_bar(width = 1, color = "black") +
  coord_polar("y") +
  labs(fill = "Experience Level", title =
  "Distribution of Individual's
  Experience Level In Data Science") +
  theme_minimal() +
  theme(legend.position = "right",
        plot.title = element_text(size = 16, face = "bold"),
        axis.title = element_blank(),
        axis.text = element_blank(),
        panel.grid = element_blank()) +
  scale_fill_manual(values = c("#56B4E9", "#E69F00", "#009E73", "#A020F0")) +
  guides(fill = guide_legend(override.aes = list(size = 3))) +
  guides(fill = guide_legend(override.aes = list(shape = 21))) +
  annotate("text", x = 1, y = 0, label = levels(salary_data$experience_level),
           color = "black", size = 4, vjust = 0.5, fill = "white") +
  geom_text(aes(label = ..count..), stat = "count", position = position_stack(vjust = 0.5),
            color = "black", size = 4)

```

Diagram 3.2.1.0.2: *Rstudio Script using ggplot to create the data plot*

The pie chart shows the distribution of individuals' experience levels in the field of data science. The chart is divided into four segments, representing different experience levels: Entry Level (EN), Executive Level (EX), Intermediate/Middle Level (MI), and Senior Level (SE). Each segment is filled with a different color.

From the pie chart, we can observe the relative proportions of individuals in each experience level category. The size of each segment indicates the frequencies of individuals at that experience level. By looking at the chart, we can see the distribution of experience levels within the dataset and get an idea of the composition of individuals in different levels of expertise in data science.

3.2.1.1 Visualization and Presentation of the Chart

The provided chart is a categorical pie chart representing the distribution of individuals' experience levels in the field of data science. Explanation for each of the elements used to generate the pie chart is as below:

Colors

The `scale_fill_manual()` function is used to manually specify the fill colors for the different experience levels. Four colors are chosen in hexadecimal value which are "#56B4E9", "#E69F00", "#009E73", and "#A020F0". These colors are visually distinct and help differentiate between the experience levels in the chart.

Title

The chart title, "Distribution of Individual's Experience Level in Data Science," provides a clear indication of the chart's purpose. It highlights the focus on analyzing the experience levels of individuals in the data science field.

Theme

The `theme_minimal()` function is used to apply a minimalistic theme to the chart. This removes unnecessary gridlines and axis text, keeping the focus on the essential elements of the plot.

Legend and Labels

The `legend.position = "right"` code places the legend on the right side of the chart. It provides a key to understanding the fill colors used for each experience level. The `override.aes` arguments in the `guides()` function control the size and shape of the legend symbols, ensuring they match the filled segments of the pie chart. There are 4 experience levels of individuals labelled in short phrases which are EN: Entry Level, EX: Executive Level, MI: Middle / Intermediate Level and SE: Senior Level which represent the highest level of experience among the four of them.

The `annotate()` function adds text labels inside the pie chart, representing the different experience levels for each segments. These labels are positioned at the center of each segment and are sized appropriately for readability. The `geom_text()` function displays the count of individuals for each experience level within the chart's segments.

3.2.1.2 Findings and Insights From the Analysis

The categorical pie chart provides insights into the distribution of experience levels among individuals in the data science field. There are few key findings and their interpretation from the result of the chart.

Firstly, the analysis highlights that a significant proportion of data science professionals fall into the "Intermediate" experience level category (MI), indicating a moderate level of experience. Aspiring data scientists can consider this finding as an indication of the typical career progression within the education category. It emphasizes the importance of acquiring foundational knowledge and gaining practical experience to move from entry-level positions to intermediate roles in the data science field.

Next, the pie chart provides valuable information for data science educators in designing educational programs that can cater to students at different experience levels. It shows that individuals at various stages, including entry-level (EN), intermediate (MI), and senior-level (SE), are part of the data science workforce. Educators can focus on developing curricular that will offer comprehensive coverage of concepts, tools, and techniques for each experience level, ensuring a well-rounded education for their students in data science education sector.

Then, the distribution of experience levels offers insights into the evolution of the data science field within the education category. Researchers and academics can study the career progression of professionals by analyzing the proportion of individuals at each experience level. It helps identify trends, such as the increasing number of intermediate-level professionals over time, and enables further investigation into the factors contributing to their growth in the data science field.

Next, the pie chart can assist hiring managers and recruiters in understanding the talent pool available in the education category. It indicates the presence of individuals at different experience levels, including entry-level (EN), intermediate (MI), and executive-level (EX). Hiring managers can align their recruitment strategies with the identified distribution to attract suitable candidates for various job roles.

Lastly, policymakers interested in the data science field within the education sector can utilize the pie chart to assess the distribution of experience levels among professionals. It can provide an overview of the workforce composition, indicating the presence of both entry-level and experienced individuals. Policymakers can use this information to develop policies that promote skill development, training opportunities, and educational initiatives tailored to each experience level, ensuring a sustainable and thriving data science ecosystem in the country.

In summary, the categorical pie chart reveals the distribution of experience levels in the data science field. It highlights the typical career progression, provides insights for educational program design, facilitates research on industry dynamics, aids in talent acquisition, and guides policymaking efforts within the education sector of data science.

3.2.1.3 Suggestion For the Findings

Based on the findings obtained from the visualization of the distribution of experience levels in the data science field, the following suggestions can be made.

Skill Development Programs

Firstly, aspiring data scientists can benefit from skill development programs that cater to different experience levels. Institutions and organizations offering data science education can use the insight to improve the design comprehensive courses and training programs that cover foundational concepts for entry-level individuals and more advanced topics for intermediate and senior-level professionals. This can provide a well-rounded education and continuous professional growth.

Career Guidance

Data science educators can provide career guidance and counseling services to help students navigate the different experience levels in the field. By offering insights into the typical career progression and opportunities at each level, educators can assist students in making informed decisions about their educational paths and career goals in data science field.

Research on Career Trajectories

Researchers and academics can conduct studies to further explore the factors influencing career trajectories in the data science field. By analyzing the distribution of experience levels over time, researchers can identify trends, such as the demand for specific skills or the emergence of new roles. This research can inform the development of targeted educational programs and industry-oriented curriculum updates.

Policy Support

Policymakers in the education sector can use the findings to shape policies that foster the growth of the data science field. This can include allocating resources for training programs, promoting collaborations between academia and industry, and facilitating internships or apprenticeships to bridge the gap between education and employment. Policies that support continuous professional development and create pathways for career progression can further strengthen the data science ecosystem in certain country.

By implementing these suggestions, the aspiring data analysts, data science educator, researcher and academics in data science and policymakers can ensure a well-aligned and robust data science workforce that meets the evolving needs of the industry and promotes career advancement opportunities for individuals at different experience levels.

3.2.2 Continuos Category: Distribution of Remote Workplace Ratio of Employees

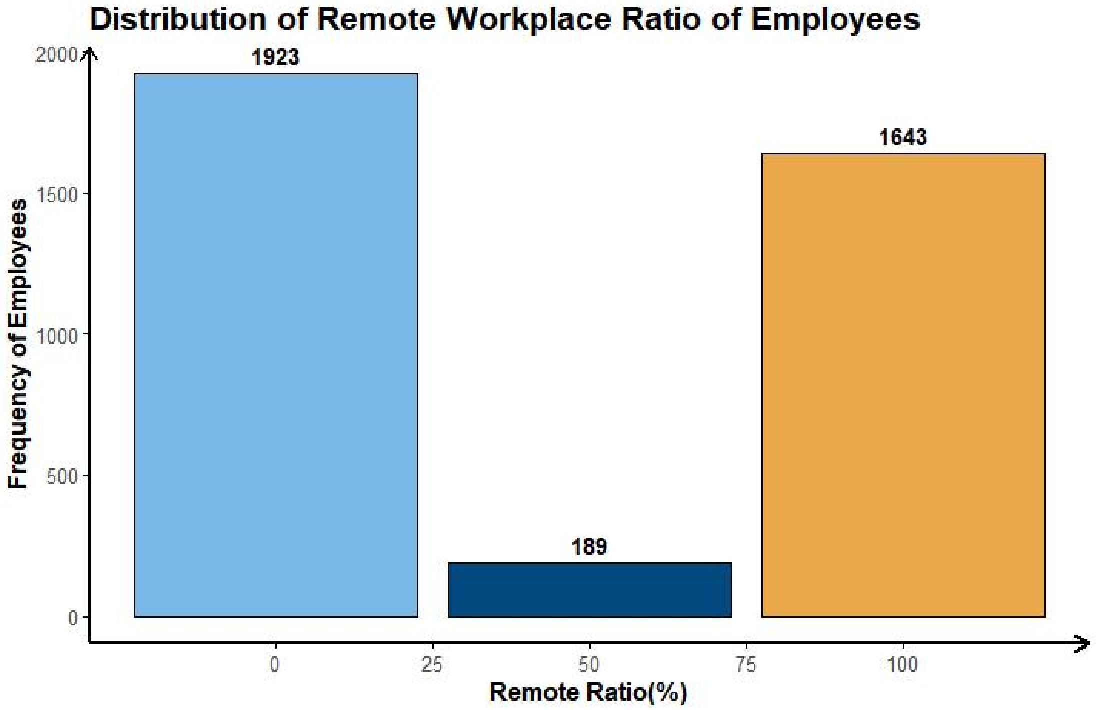
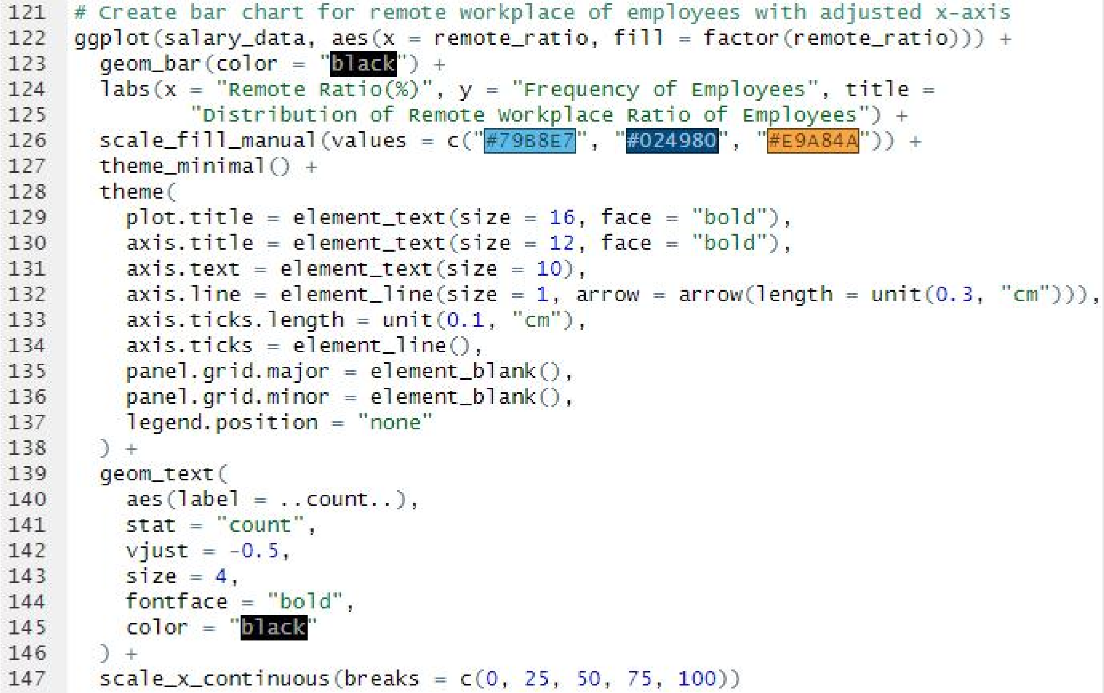


Diagram 3.2.2.0.1: *Bar Chart : Distribution of Remote Workplace Ratio of Employees*

Rstudio Script



```
# Create bar chart for remote workplace of employees with adjusted x-axis
ggplot(salary_data, aes(x = remote_ratio, fill = factor(remote_ratio))) +
  geom_bar(color = "black") +
  labs(x = "Remote Ratio(%)", y = "Frequency of Employees", title =
         "Distribution of Remote Workplace Ratio of Employees") +
  scale_fill_manual(values = c("#79B8E7", "#024980", "#E9A84A")) +
  theme_minimal() +
  theme(
    plot.title = element_text(size = 16, face = "bold"),
    axis.title = element_text(size = 12, face = "bold"),
    axis.text = element_text(size = 10),
    axis.line = element_line(size = 1, arrow = arrow(length = unit(0.3, "cm"))),
    axis.ticks.length = unit(0.1, "cm"),
    axis.ticks = element_line(),
    panel.grid.major = element_blank(),
    panel.grid.minor = element_blank(),
    legend.position = "none"
  ) +
  geom_text(
    aes(label = ..count..),
    stat = "count",
    vjust = -0.5,
    size = 4,
    fontface = "bold",
    color = "black"
  ) +
  scale_x_continuous(breaks = c(0, 25, 50, 75, 100))
```

Diagram 3.2.2.0.2: *Rstudio Script using ggplot to create the data plot*

The provided bar chart represents the distribution of remote workplace ratios among employees in the data science field. The chart showcases the frequencies of different remote workplace ratios, represented on the x-axis as percentages (%), and the y-axis represents the frequency of employees in each remote ratio category. The chart title, "Distribution of Remote Workplace Ratio of Employees," clearly conveys the purpose of the visualization.

Next, from the bar chart, we can see that most of the employees are doing their work fully at their workplace(0%) with frequency of 1923, followed by doing full work remotely from their workplace(100%) with frequency of 1643 and lastly half work done remotely and in their workplace(50%) with frequency of 189.

3.2.2.1 Visualization and Presentation of the Chart

This is a bar chart to represent the distribution of remote workplace ratios among employees. The chart visualizes the frequencies of different "Remote Ratio" categories and incorporates various design elements to enhance its clarity and effectiveness. The elements that are presented in the chart are going to be explained such as the following.

Colors and Contrast

The bar chart uses a hexadecimal value color palette with three distinct colors, namely "#79B8E7", "#024980", and "#E9A84A". These colors are used to represent different categories of the "Remote Ratio" variable, providing a visual distinction between the bars. The contrasting colors enhance the readability and make it easier to differentiate between the bars.

Size

The bars in the bar chart have a consistent width, defined by the width parameter set to its default value of 1. This ensures that each bar is equally visible and contributes to the overall representation of the distribution for the workplace remote ratio.

Theme

The bar chart is designed with a minimal theme using theme_minimal(). This results in a clean and uncluttered visual appearance, allowing the focus to be on the data itself. The absence of unnecessary background grids and borders in the minimal theme helps in emphasizing the main elements of the chart.

Scale

The fill scale is manually defined using scale_fill_manual(). It assigns the specified color values to the factor levels of the "Remote Ratio" variable. This ensures that each category consistently retains its assigned color throughout the chart. The defined color palette enables quick and intuitive recognition of different categories of the x and y-axis variables.

Shape

The chart primarily consists of rectangular bars, which are the default shape for a bar chart. The use of bars facilitates the comparison of frequencies or counts associated with each "Remote Ratio" category. The rectangular shape is ideal for representing continuous data on a discrete scale.

Label and Legend

The chart includes a title, axis labels, and a count label for each bar. The title, "Distribution of Remote Workplace Ratio of Employees," provides an overview of the chart's purpose. The x-axis label, "Remote Ratio (%)," describes the continuous variable being represented. The y-axis label, "Frequency of Employees," indicates the quantity that are being measured. The count labels on top of each bar displayed the frequency of employees corresponding to each "Remote Ratio" category. The absence of a legend is intentional as the fill colors and categories are directly used and labeled within the chart.

3.2.2.2 Findings and Insights from the analysis

The findings of this chart have implications for various target audiences within the education category.

Firstly, aspiring data scientists can gain insights into the prevailing work arrangements in the industry, helping them plan their career paths and make informed decisions regarding remote work opportunities.

Next, data science educators can utilize the chart to discuss the evolving nature of work arrangements in the field, while researchers and academics can analyze the distribution to explore the impact of remote work on productivity, job satisfaction, and overall performance within the data science domain.

Afterwards, for hiring managers and recruiters, the chart provides valuable insights into the expectations and preferences of data science professionals regarding remote work. This information can guide hiring decisions and help tailor job offerings to attract top talent in the data science field.

Lastly, policymakers interested in labor market dynamics and the future of work can examine the chart to understand the prevalence of remote work in the data science sector. Such insights can inform the formulation of policies that support flexible work arrangements and foster growth in data sience sector of the country.

Overall, the bar chart offers a visual representation of the distribution of remote workplace ratios among employees in the data science field, specifically catering to educate and provide useful insights to the target audience. It provides valuable insights to aspiring data scientists, data science educators, researchers, hiring managers, recruiters, and policymakers, enabling them to make informed decisions, understand prevailing work dynamics, and support the growth of the education sector within data science.

3.2.2.3 Suggestion from the Findings

Based on the findings obtained from the visualization of the distribution of remote workplace ratios among employees in the data science field, the following suggestions can be made:

Embrace Flexible Work Policies

Employers in the data science field should consider implementing flexible work policies that allow employees to work remotely to some extent. The chart shows a significant portion of data science professionals that have the opportunity to work partially or fully remotely. By embracing flexible work arrangements, organizations can attract top talent and thus provide a conducive work environment that promotes work-life balance.

Invest in Remote Collaboration Tools

As remote work becomes more prevalent, it is crucial for organizations to invest in robust remote collaboration tools and technologies. This will enable seamless communication, collaboration, and knowledge sharing among remote teams from the organization or company. Employers should ensure that their remote workforce has access to reliable and useful tools and platforms that can facilitate virtual meetings, project management, and data sharing.

Provide Remote Work Training and Support

To ensure the success of remote work arrangements, organizations should offer training and support to employees. This can include guidance on effective remote work practices, time management strategies, and virtual communication skills.

By equipping employees with the necessary skills and resources, employers can optimize productivity and maintain a good remote work culture.

Overall, the findings from the visualization suggest that remote work is prevalent in the data science field. Employers and organizations can leverage this trend by embracing flexible work policies investing in remote collaboration tools and providing training and support.. By doing so, they can adapt to evolving work preferences, attract and retain top talent, and create a productive and engaging work environment in the education category of the data science field.

3.3 Bivariate Analysis Chart

3.3.1 Line Chart: Average Salary by Work Year and Company Size

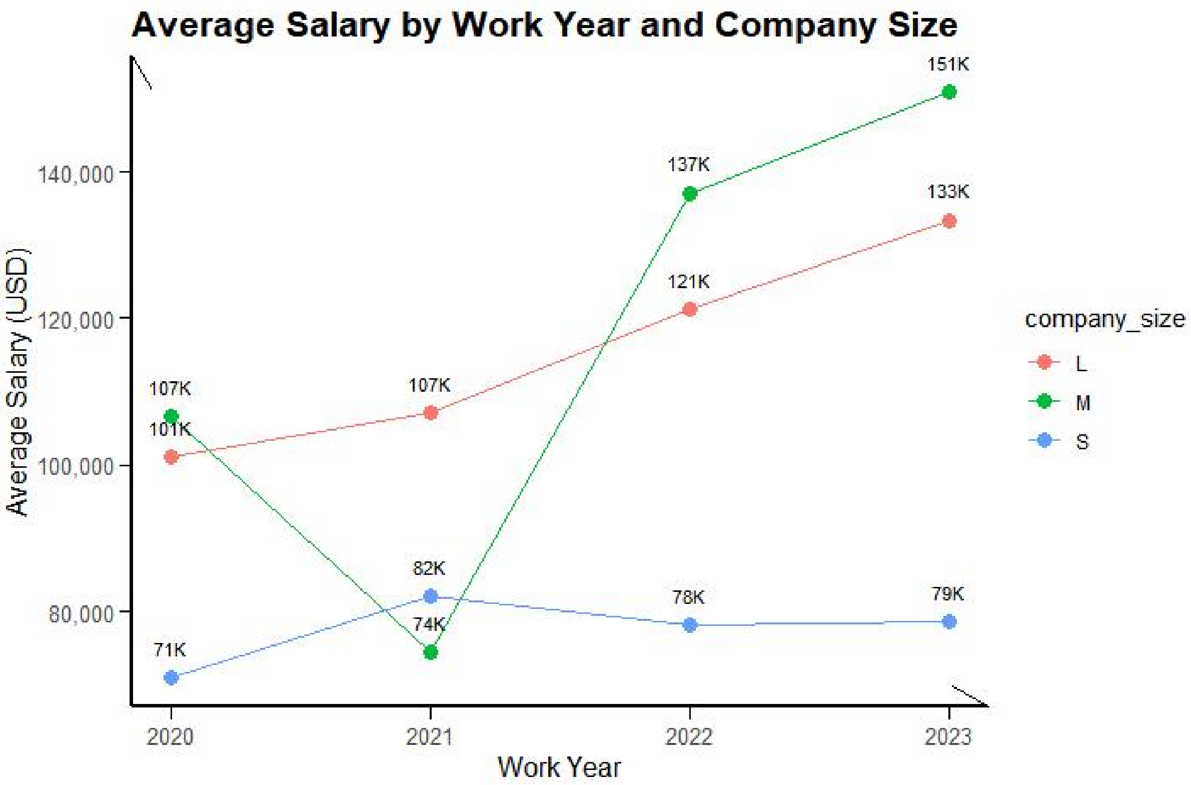
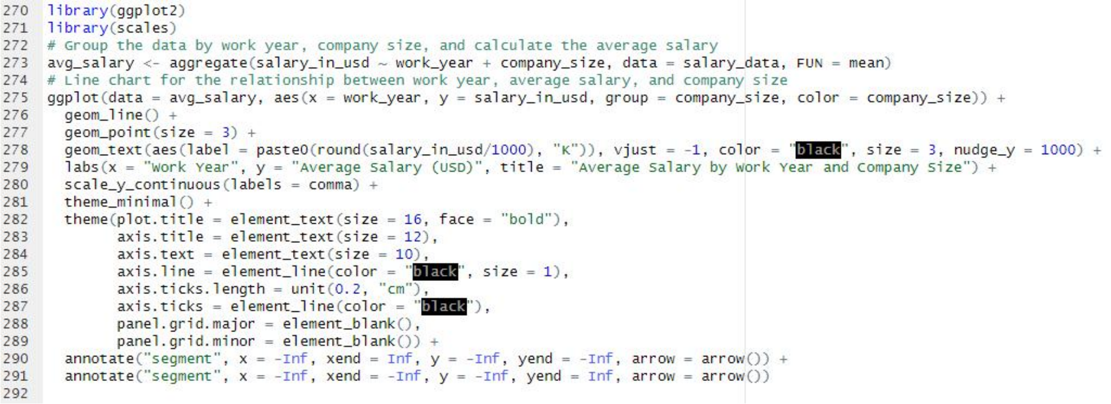


Diagram 3.3.1.0.1: *Line Chart of Average Salary by Work Year and Company Size*

Rstudio Script



```
library(ggplot2)
library(scales)
# Group the data by work year, company size, and calculate the average salary
avg_salary <- aggregate(salary_in_usd ~ work_year + company_size, data = salary_data, FUN = mean)
# Line chart for the relationship between work year, average salary, and company size
ggplot(data = avg_salary, aes(x = work_year, y = salary_in_usd, group = company_size, color = company_size)) +
  geom_line() +
  geom_point(size = 3) +
  geom_text(aes(label = paste0(round(salary_in_usd/1000), "K")), vjust = -1, color = "black", size = 3, nudge_y = 1000) +
  labs(x = "Work Year", y = "Average Salary (USD)", title = "Average Salary by Work Year and Company Size") +
  scale_y_continuous(labels = comma) +
  theme_minimal() +
  theme(plot.title = element_text(size = 16, face = "bold"),
        axis.title = element_text(size = 12),
        axis.text = element_text(size = 10),
        axis.line = element_line(color = "black", size = 1),
        axis.ticks.length = unit(0.2, "cm"),
        axis.ticks = element_line(color = "black"),
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank()) +
  annotate("segment", x = -Inf, xend = Inf, y = -Inf, yend = -Inf, arrow = arrow()) +
  annotate("segment", x = -Inf, xend = -Inf, y = -Inf, yend = Inf, arrow = arrow())
```

Diagram 3.3.1.0.2: *Rstudio Script using ggplot to create the data plot*

The line chart displays the relationship between work years, average salary, and company size in the dataset. The chart helps visualize how average salary varies with work years and different company sizes.

Each line on the chart represents a specific company size category. The x-axis represents the number of work years, indicating the level of professional experience. The y-axis represents the average salary in USD. Firstly,the lines show the trend of average salary changes with increasing work years. You can see whether the salaries generally increase, decrease, or remain constant as work years increase.

Then, the different-colored lines represent the different company sizes. By comparing the lines, you can understand how average salaries differ across company size. For example, you can observe if larger companies tend to offer higher salaries compared to smaller ones. Next, the points on the lines represent the average salary values. The associated labels indicate the average salary in thousands (e.g., "50K" represents $50,000). These labels help provide a sense of the salary mean for each data point.

3.3.1.1 Visualization and Presentation of the Chart

The provided code generates a line chart that displays the relationship between work years, average salary, and company size. The elements of visualization are as the following.

Colors and Contrast

The colors in the chart are determined by the `color` aesthetic, which assigns different colors to each company size category. Each line and associated points are colored accordingly, allowing for visual differentiation between different company sizes. The use of contrasting colors ensures clarity and readability of the chart.

Size

The `size` parameter in `geom_point()` determines the size of the points representing each data point on the chart. In this case, the size is set to 3, indicating that the points will be relatively large and more prominent in the chart.

Theme

The chart is styled using the `theme_minimal()` function, which provides a clean and minimalist appearance with a white background. Additional theme settings modify various elements, such as the plot title (`plot.title`), axis titles (`axis.title`), axis labels (`axis.text`), axis lines (`axis.line`), tick lengths (`axis.ticks.length`), and grid lines (`panel.grid.major` and `panel.grid.minor`). These settings contribute to the overall visual style and readability of the chart.

Scale

The `scale_y_continuous()` function is used to format the y-axis labels. The `labels = comma` argument applies comma formatting to the y-axis labels, making them more readable by separating thousands with commas.

Label and Legend

The `geom_text()` function adds text labels to the chart. The `label` parameter specifies the text to be displayed, which, in this case, shows the average salary values in thousands (rounded to the nearest whole number). The legend is automatically generated based on the `color` aesthetic, representing different company sizes. The labels and legend provide additional information about the data and assist in understanding the chart.

3.3.1.2 Findings and Insights From the Analysis

The line chart illustrates the relationship between work years, average salary, and company size in the data science field. The chart is based on aggregated data, grouping the information by work year and company size and calculating the average salary for each combination. The key findings and intepretation is as below.

Firstly, the line chart shows how the average salary in USD changes with increasing work years in the data science field. As the work years increase, the average salary tends to rise as well. This finding suggests that experience plays a significant role in salary progression within the data science industry. Aspiring data scientists can use this information to understand the potential financial growth over their career trajectory and set realistic salary expectations when applying for any position in data science field.

Next, the line chart differentiates the lines by company size, providing insights into how company size influences average salaries in the data science field. By examining the trends of different lines, we can observe if larger or smaller companies offer higher average salaries. Furthermore, this information can be valuable for data science educators, recruiters, and policymakers. Educators can emphasize the potential salary benefits associated with joining companies of different sizes to guide students' career decisions. Recruiters can use this insight to position job opportunities and compensation packages based on company size. Policymakers can consider these salary trends when formulating policies to promote the growth of data science industries within different company size.

Then, by examining the intersections of the lines representing different company sizes, we can compare the average salaries for different work years within each company size category. This analysis allows for a more nuanced understanding of how salaries vary based on both work experience and company size. To prove the point, researchers and academics can explore these patterns to study factors influencing salary discrepancies, such as industry maturity, market competition, or resource allocation across companies of varying sizes.

In summary, the line chart provides insights into the relationship between work years, average salary, and company size in the data science field. The findings can inform aspiring data scientists about salary growth potential, assist data science educators and recruiters in guiding career decisions, and offer insights for policymakers aiming to develop effective policies to support the data science industry for the growth of the country and the people.

3.3.1.3 Suggestion From the Findings

Based on the findings obtained from the visualization, the following suggestions can be made.

Emphasize the Importance of Experience

The relationship between work years and average salary highlights the significance of gaining experience in the data science field. Aspiring data scientists should focus on building their skills and expertise through practical projects, internships, and industry experience. Furthermore, they can consider seeking opportunities to work with companies that offer growth prospects and mentorship programs to accelerate their career progression and increase their earning potential in the data science field.

Consider the Impact of Company Size

The variation in average salaries across different company sizes indicates the influence of company size on compensation. Data science professionals can evaluate their preferences regarding company size based on their salary expectations. Larger companies may offer higher salaries but could have more competition and a more structured career path, while smaller companies may provide more opportunities for growth and a greater level of autonomy. It is essential for individuals to weigh these factors and align their career goals with the company size that aligns with their aspirations.

Negotiation and Market Research

Understanding the average salaries at different work years and company sizes empowers data science professionals during salary negotiation. Armed with this knowledge, they can benchmark their expected compensation against industry standards and make informed decisions when negotiating job offers or salary increases. To conclude, they can conduct thorough market research, leveraging salary surveys, and networking with professionals in the field can provide additional insights into salary expectations and trends.

Continuous Learning and Skill Development

The correlation between work years and average salary underscores the importance of continuous learning and skill development. Data science professionals should invest in ongoing education, stay updated with the latest tools and techniques, and develop expertise in emerging areas such as artificial intelligence, machine learning, or big data analytics. By enhancing their skills, professionals can position themselves for higher-paying roles and remain competitive in the rapidly evolving field of data science.

Overall, the findings from the visualization provide insights that can guide data science professionals in their career decisions, salary negotiations, and professional development. By considering these suggestions, individuals can navigate the data science landscape more effectively and make informed choices that align with their goals in the sector.

3.3.2 Grouped Bar Chart: Job Title Count by Work Year and Experience Level

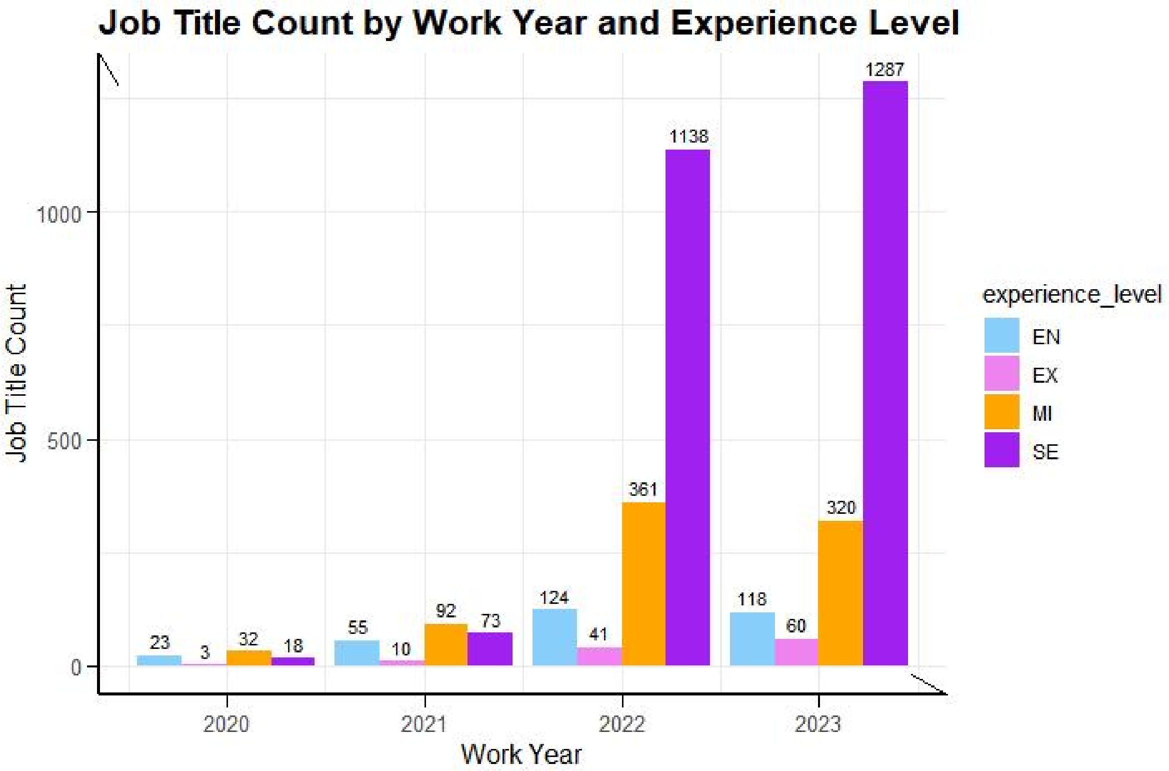
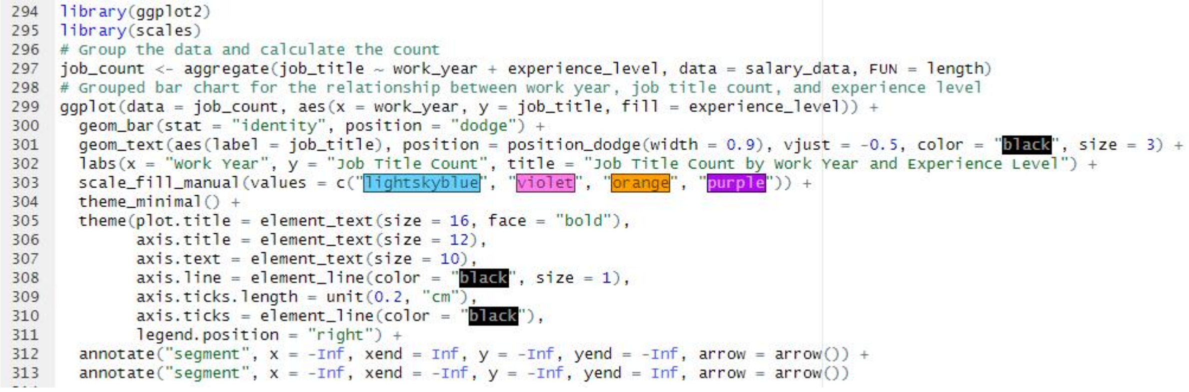


Diagram 3.3.2.0.1: *Grouped Bar Chart of Job Title Count by Work Year and Experience Level*

Rstudio Script



```
library(ggplot2)
library(scales)
# Group the data and calculate the count
job_count <- aggregate(job_title ~ work_year + experience_level, data = salary_data, FUN = length)
# Grouped bar chart for the relationship between work year, job title count, and experience level
ggplot(data = job_count, aes(x = work_year, y = job_title, fill = experience_level)) +
  geom_bar(stat = "identity", position = "dodge") +
  geom_text(aes(label = job_title), position = position_dodge(width = 0.9), vjust = -0.5, color = "black", size = 3) +
  labs(x = "Work Year", y = "Job Title Count", title = "Job Title Count by Work Year and Experience Level") +
  scale_fill_manual(values = c("lightskyblue", "violet", "orange", "purple")) +
  theme_minimal() +
  theme(plot.title = element_text(size = 16, face = "bold"),
        axis.title = element_text(size = 12),
        axis.text = element_text(size = 10),
        axis.line = element_line(color = "black", size = 1),
        axis.ticks.length = unit(0.2, "cm"),
        axis.ticks = element_line(color = "black"),
        legend.position = "right") +
  annotate("segment", x = -Inf, xend = Inf, y = -Inf, yend = -Inf, arrow = arrow()) +
  annotate("segment", x = -Inf, xend = -Inf, y = -Inf, yend = Inf, arrow = arrow())
```

Diagram 3.3.2.0.2: *Rstudio Script using ggplot2 to create the data plot*

The grouped bar chart shows the relationship between work years, job title count, and experience level in the dataset. The chart helps visualize the distribution of job titles across different work years and experience levels.

The x-axis represents the number of work years, indicating the level of professional experience. Then, the y-axis represents the count of job titles, showing the number of individuals holding each job title category within each work year.

Next, the bars are grouped by different experience levels, which are represented by different colors. Each group shows the count of job titles for a specific experience level within each work year. By examining the chart, we see that the grouped bars provide a visual comparison of job title counts across different experience levels for each work year. We can observe how job titles are distributed among different experience levels.

Furthermore, by comparing the bars within each work year, we can identify any shifts or trends in job title counts across experience levels. This can help us understand how the distribution of job titles changes as professionals gain more work experience. The different colors of the bars represent different experience levels. We can compare the heights of the bars within each work year to understand the relative count of job titles for each experience level.

3.3.2.1 Visualization and Presentation of the Chart

The grouped bar chart can provide insights into the distribution of job titles across different work years and experience levels according to its visualization and presentation. The elements within the visualization is as the following.

Colors and Contrast

The `fill` aesthetic within the `aes()` function determines the colors of the bars in the grouped bar chart. The `scale_fill_manual()` function is used to manually specify the colors for each experience level. In this case, the colors are set to "lightskyblue", "violet", "orange", and "purple". These color choices create contrast and differentiate the bars for each experience level according to the dataset.

Size

The size of the text labels on the bars is controlled by the `size` parameter in the `geom_text()` function. In this case, the size is set to 3, indicating that the text labels should be relatively small.

Theme

The appearance of the plot is controlled by the `theme_minimal()` function, which sets a minimalistic theme with a white background. Additional theme settings modify various elements, such as the plot title (`plot.title`), axis titles (`axis.title`), axis labels (`axis.text`), axis lines (`axis.line`), tick lengths (`axis.ticks.length`), and grid lines (`panel.grid.major` and `panel.grid.minor`). These provide overall visual style of the plot.

Scale

The `scale_fill_manual()` function is used to manually define the colors for the fill aesthetic, allowing customization of the legend colors. By specifying the colors using the `values` parameter, the function assigns the desired colors to each level of the `experience_level` variable of the chart.

Label and Legend

The `geom_text()` function is used to add text labels to the bars in the plot. The `label` parameter specifies the text to be displayed, which, in this case, is the `job_title` count. The `geom_text()` function is positioned using `position_dodge()` with a width of 0.9, ensuring that the text labels appear slightly offset from the bars for clarity.

Next, the `labs()` function is used to set the labels for the x-axis, y-axis, and the plot title. The legend is automatically generated based on the `fill` aesthetic, indicating the relationship between the colors and the `experience_level` variable. The `legend.position` parameter within the `theme()` function is set to "right" to position the legend on the right side of the chart.

3.3.2.2 Findings and Insights from the analysis

The grouped bar chart illustrates the relationship between work years, job title count, and experience level in the data science field. From the chart, we can derive several important findings and interpretations.

Firstly, the chart provides valuable insights for individuals aspiring to pursue a career in data science. It demonstrates the distribution of job titles across different experience levels and work years. Furthermore, aspiring data scientists can use this information to understand the demand for specific job titles and the typical career progression within the field. It can guide their decision-making process in terms of skill development and goal settings.

Next, the visualization offers meaningful information for data science educators. By analyzing the count of job titles at different experience levels and work years, educators can identify the skills and knowledge areas that are in high demand in the industry. This knowledge can inform curriculum development and help educators tailor their programs to meet the needs of students and the evolving job market of data science field.

Then, the chart provides researchers and academics with insights into the job landscape within the data science field. By examining the count of job titles across experience levels and work years, researchers can identify trends and patterns in the industry. This information can guide further research on topics such as career progression, job satisfaction, and the impact of experience on job opportunities within the field of data science for related topics.

Afterwards, the visualization offers valuable information for hiring managers and recruiters in the data science industry. By analyzing the count of job titles across experience levels and work years, they can gain a better understanding of the talent pool and the availability of candidates for different positions. This knowledge can inform their recruitment strategies, job postings, and improve candidate selection processes.

Lastly, policymakers interested in the data science field can utilize the chart to gain insights into the job market dynamics. It can provide information on the distribution of job titles across experience levels and work years, helping policymakers understand the demand for specific skills and the potential gaps in the workforce. This knowledge can guide policies related to education, training, and workforce development initiatives to boost data science industry in any country.

In summary, the grouped bar chart provides valuable insights into the relationship between work years, job title count, and experience level in the data science field. It offers meaningful information for aspiring data scientists, data science educators, researchers and academics, hiring managers and recruiters, and policymakers, thus enable them to make informed decisions and strategies in their respective domains.

3.3.2.3 Suggestion From the Findings

Based on the findings obtained from the visualization, the following suggestions can be made.

Skill Development Programs

As the data science field continues to evolve, it is crucial to focus on skill development programs that cater to the demand for specific job titles. Educators and training institutions can align their curriculum and courses to address the skills needed at different experience levels and work years. This include offering specialized training in areas where job titles are in high demand, such as machine learning, data engineering, or data visualization.

Career Path Guidance

The distribution of job titles across experience levels and work years provides valuable insights for individuals looking to progress in their data science careers. Career guidance resources and mentorship programs can be developed to help aspiring data scientists understand the typical career progression and the skills required to advance to higher-level job titles. By providing clear pathways, the guidance can support professionals in making informed decisions about their career development in data science field.

Continuous Learning and Upskilling

Professionals in the data science field should recognize the importance of continuous learning and upskilling to remain competitive. The chart highlights the demand for job titles at different experience levels, indicating the need for ongoing professional development. Data scientists should proactively seek opportunities to enhance their skills and stay updated with the latest advancements in the field. This can include participating in workshops, attending conferences, or pursuing advanced certifications.

Overall, the findings from the visualization suggest the need for targeted skill development, career guidance and continuous learning within the data science field. By implementing these suggestions, aspiring data scientists, educators, researchers, hiring managers, and policymakers can contribute to the growth and success of the data science industry.