```
---
title: 'Unlocking Insights: Exploring Salary Dynamics and Employment Trends in Data
Science'
author: "Krishna Patel"
date: "February 25, 2024"
output:
  html_document:
    code_download: yes
    fig_caption: yes
    theme: lumen
    toc: yes
    toc_depth: 2
    df_print: kable
    toc_float:
      collapsed: no
  pdf_document:
    toc: yes
    toc_depth: '2'
---
```{r, message=FALSE}
# Required packages for our course. Do not delete.
library(tidyverse)
library(mosaic)
library(dplyr)
library(ggplot2)
library(plotly)
library(treemap)
library(devtools)
library(htmlwidgets)
library(d3treeR)
```


Introduction
The field of data science is rapidly evolving, with professionals playing pivotal roles in shaping industries across the globe. Understanding the dynamics of job roles, salaries, and related factors is crucial for both aspiring data scientists and organizations seeking to attract and retain talent. In this report, we delve into insights derived from the comprehensive dataset titled "Jobs and Salaries in Data Science," obtained from Kaggle.

Our analysis centers around several key questions aimed at uncovering patterns and trends within the data science domain:

1. **Correlation between Employee Residence and Salary Levels:** Are there discernible correlations between where data science professionals reside and the levels of their salaries? How does this correlation vary across different regions or countries?

2. **Prevalence of Job Titles Across Regions:** Do certain job titles dominate specific regions or countries? By exploring this question, we aim to discern any geographical preferences or trends in job roles within the data science field.

3. **Distribution of Work Experience Levels:** What is the distribution of respondents based on their years of work experience? Understanding this distribution provides valuable insights into the experience levels prevalent within the data science workforce.

4. **Trends in Work Experience Levels Across Locations:** Are there notable trends in work experience levels among data science professionals residing in different locations? By analyzing this aspect, we aim to identify any geographical variations in experience levels.

5. **Trend in Salaries Across Work Years:** Is there a discernible change in the average salary across different work years within the data science field?

6. **Trend in Salaries Across Experience Levels:** Is there a discernible change in the
```

average salary across different experience levels within the data science field?

Dataset Description

The "Jobs and Salaries in Data Science" dataset is a comprehensive compilation of information pertaining to salaries and related factors within the data science field. It includes details such as job title, job category, salary in various currencies, employee residence, experience level, employment type, work setting, company location, and company size. This rich dataset offers an invaluable resource for analyzing salary trends, comparing salaries across roles and regions, and understanding the factors influencing salary structures within the data industry.

In the subsequent sections of this report, we delve deeper into each research question, presenting our findings and interpretations derived from the analysis of the dataset. Through this exploration, we aim to provide actionable insights that can inform strategic decisions for both individuals and organizations operating within the dynamic landscape of data science.

Methodology

```{r}
# import the dataset
jobs_in_data <- read.csv("~/Downloads/github/salarytrend_datascience/jobs_in_data.csv")
# filter to include 25 random countries out of 83
set.seed(123)
selected_data <- jobs_in_data[, c("work_year", "job_title", "salary_in_usd",
"experience_level", "employee_residence")]
# Extract unique countries from the "employee_residence" column
unique_countries <- unique(selected_data$employee_residence)

# Select 15 random unique countries
random_15_countries <- sample(unique_countries, 25)
filtered_data <- selected_data[selected_data$employee_residence %in% random_15_countries,
]
```

Analysis

Scatterplot of Salary in USD by Employee Residence

```{r}
# Create the ggplot object
scatterplot <- ggplot(filtered_data, aes(x = employee_residence, y = salary_in_usd, color
= employee_residence)) +
  geom_point(alpha = 0.5) +
  labs(title = "Scatterplot of Salary in USD by Employee Residence",
       x = "Employee Residence",
       y = "Salary in USD") + scale_y_continuous(labels = scales::comma_format()) +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1))

# Convert ggplot object to a plotly object
interactive_scatterplot <- ggplotly(scatterplot)
interactive_scatterplot

```

The scatterplot provides a visual representation of the relationship between salary and employee residence, allowing for insights into salary distribution, regional disparities, and potential correlations. The salaries vary significantly across different employee residences, ranging from as low as $15,000 to as high as $323,905. Certain countries or regions seem to have higher salary ranges compared to others. For example, countries like France, Portugal, and Lithuania have a wide range of salary levels, including both high and moderate salaries. There are some instances of exceptionally high salaries, such as in France where the salary reaches $323,905, and in New Zealand with a salary of $125,000. Several factors could contribute to these salary differences, including local economic

conditions, cost of living, demand for specific skills, and industry specialization in certain regions.

Prevalence of Job Titles Across Regions

```{r}
# Create a stacked bar plot
interactive_stacked_bar <- filtered_data %>%
  group_by(employee_residence, job_title) %>%
  summarise(count = n()) %>%
  ggplot(aes(x = employee_residence, y = count, fill = job_title)) +
  geom_bar(stat = "identity") +
  labs(title = "Prevalence of Job Titles Across Regions",
       x = "Region or Country",
       y = "Count",
       fill = "Job Title") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

# Convert the ggplot object to an interactive plot
interactive_stacked_bar <- ggplotly(interactive_stacked_bar)
interactive_stacked_bar

```

`France`: Leads with a total count of 69, encompassing a wide range of job titles including data engineering, data science, and machine learning engineering.
`Portugal`: Follows closely with a total count of 27, with diverse roles in data analysis, data engineering, and machine learning engineering.
`Pakistan`: Shows a smaller presence with a total count of 6, mainly in data engineering, data science consultancy, and AI programming.
`Philippines`: Exhibits a limited presence with a total count of 4, focusing on roles such as business data analysis, data science management, and data analytics.
`Belgium, Lithuania, and Turkey`: Each have a moderate presence with counts ranging from 2 to 4, primarily in roles like data scientist, machine learning engineer, and AI scientist.
`Other countries`: Have minimal representation, with counts ranging from 1 to 2, encompassing specific job titles.

Distribution of Experience Levels

```{r}
# Create an interactive barplot using plot_ly
interactive_barplot <- plot_ly(data = filtered_data, x = ~experience_level, type = "histogram", marker = list(color = "brown")) %>%
  layout(title = "Distribution of Experience Levels",
         xaxis = list(title = "Experience Level"),
         yaxis = list(title = "Frequency"))

# display the interactive barplot
interactive_barplot

```

The distribution of experience levels among employees in the filtered dataset for 25 countries is as follows: there are 35 entry-level, 3 executive, 51 mid-level, and 43 senior-level positions. This distribution provides insight into the workforce composition and highlights the prevalence of mid-level and senior-level roles compared to entry-level and executive positions. Understanding this distribution can aid in workforce planning, talent management, and organizational development strategies.

Average Salary Over Years

```{r}
# Average salary over years
```

```
average_salary <- filtered_data %>%
  group_by(work_year) %>%
  summarise(avg_salary = mean(salary_in_usd))

# Create an interactive line plot for average salary over years
interactive_line_plot <- plot_ly(data = average_salary, x = ~work_year, y = ~avg_salary,
type = "scatter", mode = "lines+markers",
                                 line = list(color = "black")) %>%
  layout(title = "Average Salary Over Years",
         xaxis = list(title = "Year"),
         yaxis = list(title = "Average Salary (USD)"))

# display the interactive line plot
interactive_line_plot

```
The graph reveals a notable increase in the average salary over the years, reflecting potential trends in the job market or economic conditions. In 2020, the average salary was $59.04k, showing a slight decrese to $57.43k in 2021. However, a more significant jump occurred in 2022, with the average salary rising to $65.56k. This increase could signify factors such as industry growth, demand for skilled professionals, or inflationary pressures impacting compensation. The most substantial increase is observed in 2023, where the average salary surged to $81.09k, suggesting robust economic conditions, high demand for talent, or advancements in specialized skills driving up compensation levels. Overall, this analysis highlights the dynamic nature of salary trends and underscores the importance of monitoring and adapting to changing market conditions for both employers and employees.

Salary Distribution by Job Title

```{r}
# Create interactive boxplot for salary distribution according to job title
interactive_salary_boxplot <- plot_ly(filtered_data, x = ~job_title, y = ~salary_in_usd,
type = "box") %>%
  layout(title = "Salary Distribution by Job Title",
         xaxis = list(title = "Job Title"),
         yaxis = list(title = "Salary (USD)"))

# Print the interactive boxplot
interactive_salary_boxplot
```

The interactive boxplot analysis provides valuable insights into the salary distribution across different job titles. Among the job titles examined, "Head of Data," "Research Engineer," and "AI Developer" stand out for their notable variation in salary. The boxplot illustrates considerable variability in salaries within the roles, with some individuals potentially earning substantially higher or lower than the median. This variation could be attributed to factors such as years of experience, industry specialization, or additional qualifications.

Head of Data: The median salary for this position is $137,000, indicating a significant earning potential.

Research Engineer: With a median salary of $168,000, the role of a Research Engineer commands considerable compensation.

AI Developer: The median salary for AI Developers is $118,000, reflecting the demand for professionals skilled in artificial intelligence technologies.

Salary Distribution by Experience Level

```{r}
# salary distribution according to experience level
# Create interactive boxplot for salary distribution according to experience level
interactive_exp_boxplot <- plot_ly(filtered_data, x = ~experience_level, y =
```

```
~salary_in_usd, type = "box") %>%
  layout(title = "Salary Distribution by Experience Level",
         xaxis = list(title = "Experience Level"),
         yaxis = list(title = "Salary (USD)"))

# Print the interactive boxplot
interactive_exp_boxplot
```

The salary distribution analysis by experience level reveals distinct trends:

- Entry Level: Median salary is $40,000, typical for new or less experienced workers. This salary level reflects the compensation typically offered to individuals who are new to the workforce or have limited professional experience. Entry-level salaries may vary depending on factors such as industry, location, and specific job roles.

- Executive: Median salary jumps to $106,000, reflecting the high responsibilities of leadership roles. Executives typically hold top leadership positions within organizations and are responsible for strategic decision-making and overseeing company operations. The higher salary range for executives reflects the significant responsibilities and leadership roles they undertake.

- Mid Level: Median salary stands at $57,220, indicating moderate experience and responsibilities. Mid-level positions often require a moderate level of experience and expertise, with individuals assuming roles that involve greater responsibilities and specialized skills compared to entry-level positions. The salary reflects a midpoint between entry-level and senior-level compensation.

- Senior Level: Median salary rises to $76,050, reflecting extensive experience and leadership roles. Senior-level roles typically require extensive experience, specialized skills, and a track record of leadership and accomplishments. The higher compensation reflects the value placed on the expertise and contributions of senior professionals to organizations.

Overall, salaries increase significantly as professionals progress from entry to senior levels, reflecting their experience and contributions to organizations.

Treemap of Average Salary by Employee Residence and Job Title

```{r}
# Geographical salary distribution
# Group the data
grouped_data <- filtered_data %>%
  group_by(employee_residence, job_title, salary_in_usd) %>%
  summarise(count = n())

# Calculate average salary per group
average_salary <- grouped_data %>%
  group_by(job_title, employee_residence) %>%
  summarise(avg_salary = mean(salary_in_usd), .groups = 'drop')

# Create the treemap
data_science_treemap <- treemap(average_salary,
                                index = c("job_title", "employee_residence"),
                                vSize = "avg_salary",
                                title = "Treemap of Average Salary by Employee Residence
and Job Title")

# Print the treemap
# print(data_science_treemap)
d3tree(data_science_treemap, rootname = "Jobs")

```
In this analysis, we generated a treemap graph along with its interactive version, showcasing the average salary distribution according to employee residence and job titles.

The treemap visualization allowed us to explore the salary trends across different countries and job roles efficiently.

Upon examination of the graph, it was evident that certain job titles commanded higher average salaries across various countries. Specifically, roles such as ML engineer, data engineer, data scientist, finance data analyst, data analyst, and head of data consistently emerged with the highest average salaries.

Salary Trend for Data Scientists Over Work Years

```{r}
# salary of over years
# Filter the data for Data Scientists
data_scientist_salary <- filtered_data[filtered_data$job_title == "Data Scientist", ]

# Create the scatter plot
scatter_plot <- ggplot(data_scientist_salary, aes(x = work_year, y = salary_in_usd)) +
  geom_point(color = "blue", alpha = 0.5) +
  labs(title = "Salary Trend for Data Scientists Over Work Years",
       x = "Work Year",
       y = "Salary (USD)")

# Add trend line
scatter_plot_with_trend <- scatter_plot +
  geom_smooth(method = "lm", se = FALSE, color = "black", linewidth = 0.5)

# Convert to plotly object
interactive_plot <- ggplotly(scatter_plot_with_trend)

# Print the interactive plot
interactive_plot


# TO CHECK IF THE TREND LINE IS INCREASING/DECREASING
# Fit a linear model to the data
lm_model <- lm(salary_in_usd ~ work_year, data = data_scientist_salary)

# Get the coefficients of the linear model and slop
coefficients <- coef(lm_model)
slope <- coefficients[2]

# Check if the slope is positive, indicating an increase over years
if (slope > 0) {
  cat("The trend line indicates an increase in salaries over the years.")
} else if (slope == 0) {
  cat("The trend line indicates no change in salaries over the years.")
} else {
  cat("The trend line indicates a decrease in salaries over the years.")
}

```
Based on our analysis focusing on the data scientist role, which appears to have the highest representation across the selected countries, we investigated the salary trends over the years. Our examination, as depicted in the graph and confirmed by the linear model results, reveals a subtle but discernible increase in salary over the years. While the upward trajectory is evident, it's important to note that the rate of increase appears to be minimal. This observation suggests that while there is a positive trend in data scientist salaries over time, the growth rate may be relatively slow. This insight underscores the stability or modest growth in compensation for data scientists within the selected regions over the specified period.

Results

In this study, we analyzed data from 25 randomly selected countries, focusing on five key

variables: job title, employee residence, salary in USD, work experience, and work year. Our analysis comprised several graphical representations, each offering insights into different aspects of the dataset:

ChatGPT
Results

In this study, we analyzed data from 25 randomly selected countries, focusing on five key variables: job title, employee residence, salary in USD, work experience, and work year. Our analysis comprised several graphical representations, each offering insights into different aspects of the dataset:

- Scatterplot Analysis: We conducted a scatterplot analysis to visualize the relationship between employee residence and salary in USD.
- Stacked Barplot: A stacked barplot was utilized to illustrate the prevalence of various job titles across different employee residence countries.
- Barplot of Experience Levels: We created a barplot depicting the distribution of experience levels among the sampled data.
- Average Salary Over Years: A lineplot graph was generated to display the average salary trends over the years.
- Salary Distribution by Job Title: We employed a boxplot to showcase the distribution of salaries according to different job titles.
- Salary Distribution by Experience Level: Another boxplot was utilized to visualize the salary distribution across various experience levels.
- Treemap Visualization: A treemap was constructed to present the average salary distribution by employee residence and job title.
- Salary Trend Analysis for Data Scientists: Finally, we conducted a trend analysis specifically focusing on the salary trend for data scientists over the years.

These graphical representations offer valuable insights into the dataset, providing a comprehensive understanding of salary distributions, trends, and variations across different variables and over time.

Conclusions

In this comprehensive analysis, we delved into various aspects of salary distribution, workforce composition, and salary trends across different job titles, experience levels, and geographical regions. Through interactive visualizations and statistical analyses, we gained valuable insights into the dynamics of compensation within the industry.

The scatterplot depicting the relationship between salary and employee residence revealed significant variations in salary levels across different countries, underscoring regional disparities and potential factors influencing compensation. Furthermore, the stacked barplot provided a clear overview of the prevalence of different job titles across various employee residences, highlighting countries with a high concentration of specific roles.

Examining the distribution of experience levels among employees offered insights into the workforce composition, with a notable presence of mid-level and senior-level positions. This distribution reflects the maturity and expertise of the workforce, essential for understanding talent demographics and planning organizational strategies.

The analysis of average salary trends over the years unveiled a positive trajectory, with salaries experiencing a steady increase over time. This growth reflects broader economic conditions, industry demand, and evolving skill requirements contributing to salary escalations across different job roles and regions.

The examination of salary distributions by job title and experience level provided a nuanced understanding of compensation structures within the industry. Roles such as Head of Data, Research Engineer, and AI Developer emerged with notable salary variations, indicative of the diverse skillsets and responsibilities associated with these positions. Additionally, the analysis highlighted significant differences in salaries across experience levels, emphasizing the importance of expertise and seniority in driving compensation levels.

Finally, the treemap visualization offered a comprehensive view of average salary distributions across employee residences and job titles, highlighting lucrative roles and regions within the industry. By focusing on the data scientist role, we observed a subtle yet discernible increase in salary over the years, indicating stable or modest growth in compensation for professionals within this domain.

In conclusion, this analysis provides valuable insights for industry stakeholders, policymakers, and professionals seeking to understand salary dynamics, workforce trends, and regional variations within the field. By leveraging interactive visualizations and statistical analyses, we have shed light on key factors shaping compensation trends, offering a robust foundation for informed decision-making and strategic planning in the ever-evolving landscape of the industry.