Dream jobs - analysis on the job market in US

INFO 526 - Fall 2024 - Project 02: analysis on the job market in US

DreamJobs

Abstract

Introduction

The data set used for this project spans from 2020 to 2024, combining data sourced from Kaggle (2020–2023) with 2024 data scraped from job websites like Monster and Indeed. Additionally, survey responses from students in the class provided insights into job opportunities of interest, aiding in data collection. After cleaning, the data set comprises 12,419 rows and 12 columns, making it a comprehensive resource for analyzing the job market.

This project examines salary trends over the past four years, focusing on variations by job category, experience level, and company size. It also explores how work settings (remote vs. in-person) have evolved, their correlation with employee residence and salary levels, and general trends in job availability. By identifying hidden patterns, this analysis provides valuable insights into the evolving job market.

Load necessary packages

```
pacman::p_load(ggplot2, dplyr, ggrepel, gganimate, png, magick, ggridges,scales,circlize)
```

Loading the dataset

```
data <- read.csv("updated_data.csv")
```

How do salaries vary by job category across different experience levels?

Introduction: Understanding how salaries differ across job categories and experience levels provides insights into the value placed on different roles and the progression of compensation with experience. This analysis can help identify which roles offer the highest earning potential and how experience influences salary growth within each category.

Approach:

Use box plots to visualize the salary distribution across job categories, segmented by experience level.

Highlight variations in median and spread of salaries for each category.

Analyze patterns to identify high-paying job categories for entry-level, mid-level, and senior professionals.

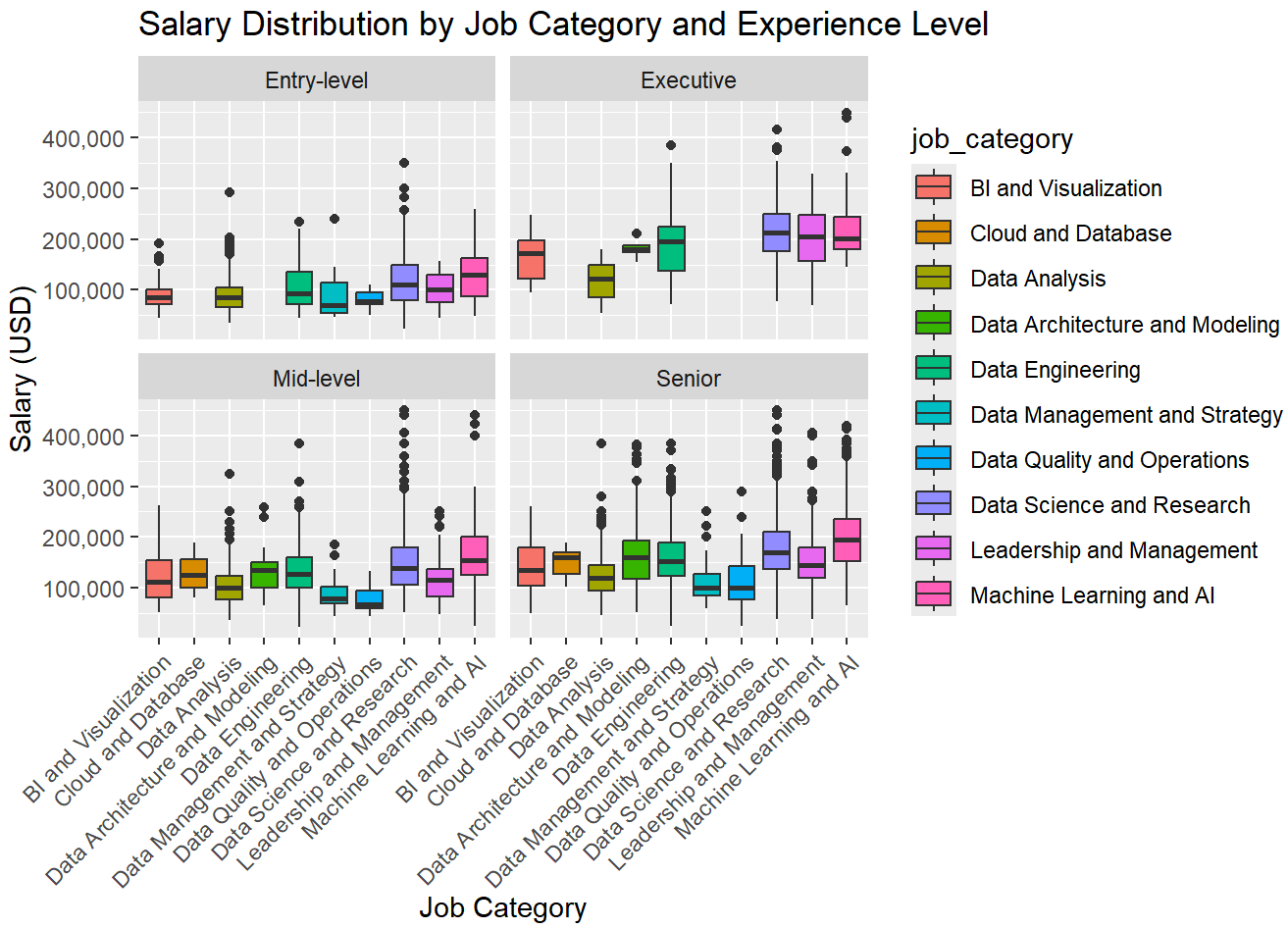
Analysis

```
# ensuring the salary is in numeric column
data$salary_in_usd <- as.numeric(data$salary_in_usd)
```

Creating the box plot with formatted y-axis

```
p <- ggplot(data, aes(x = job_category, y = salary_in_usd, fill = job_category)) +
  geom_boxplot() +
  facet_wrap(~experience_level) +
  labs(
    title = "Salary Distribution by Job Category and Experience Level",
    x = "Job Category",
    y = "Salary (USD)"
  ) +
  scale_y_continuous(labels = comma) +
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1)
  )

print(p)
```




```
ggsave("distplot.png", plot = p, width = 8, height = 6, dpi = 300)
```

This graph illustrates salary distributions across various job categories and experience levels, segmented into four groups: Entry-level, Mid-level, Senior, and Executive. The x-axis represents job categories, such as “Data Engineering,” “Data Science and Research,” and “Machine Learning and AI,” while the y-axis shows salaries in USD. Each boxplot conveys the spread of salaries within a specific job category and experience level, highlighting the median (middle line), interquartile range (the box), and outliers (dots beyond the whiskers). The job categories are color-coded for clarity, and the legend provides easy identification of each domain.

Entry-level roles tend to have lower salary ranges, with medians clustering between $100,000 and $150,000 across most categories. Some fields, like “Machine Learning and AI” and “Leadership and Management,” show slightly higher entry-level salaries, suggesting higher demand or specialization. As experience progresses to Mid-level and Senior roles, the salary ranges and medians increase notably. For instance, “Data Engineering” and “Machine Learning and AI” exhibit substantial growth, reflecting the importance of these roles in driving business value and innovation. The variability in salaries also increases at higher experience levels, as indicated by longer whiskers and more outliers.

Executive roles show the most significant salary ranges, with categories like “Leadership and Management” and “Machine Learning and AI” commanding medians close to or exceeding $300,000. These roles exhibit the widest variability, with outliers extending beyond $400,000, reflecting the diverse factors influencing executive compensation, such as company size, industry, and location. Job categories like “BI and Visualization” and “Data Quality and Operations” generally have narrower salary ranges, indicating more consistency in pay. This visualization provides insights into how specialization, demand, and experience level affect earning potential, helping to identify trends and opportunities in the data and technology sectors.

Which job titles are most commonly associated with executive-level experience?

Introduction: Identifying the most common job titles at the executive level can provide insights into the roles typically held by individuals with significant experience and leadership responsibilities. This analysis highlights the demand for certain executive roles and helps identify career paths that commonly lead to these positions.

Approach:

Filter the dataset to include only rows where the experience level is “Executive.”

Count the frequency of each job title to determine the most common executive-level roles.

Use a horizontal bar chart to visualize the frequency of job titles, making it easy to identify prominent roles.

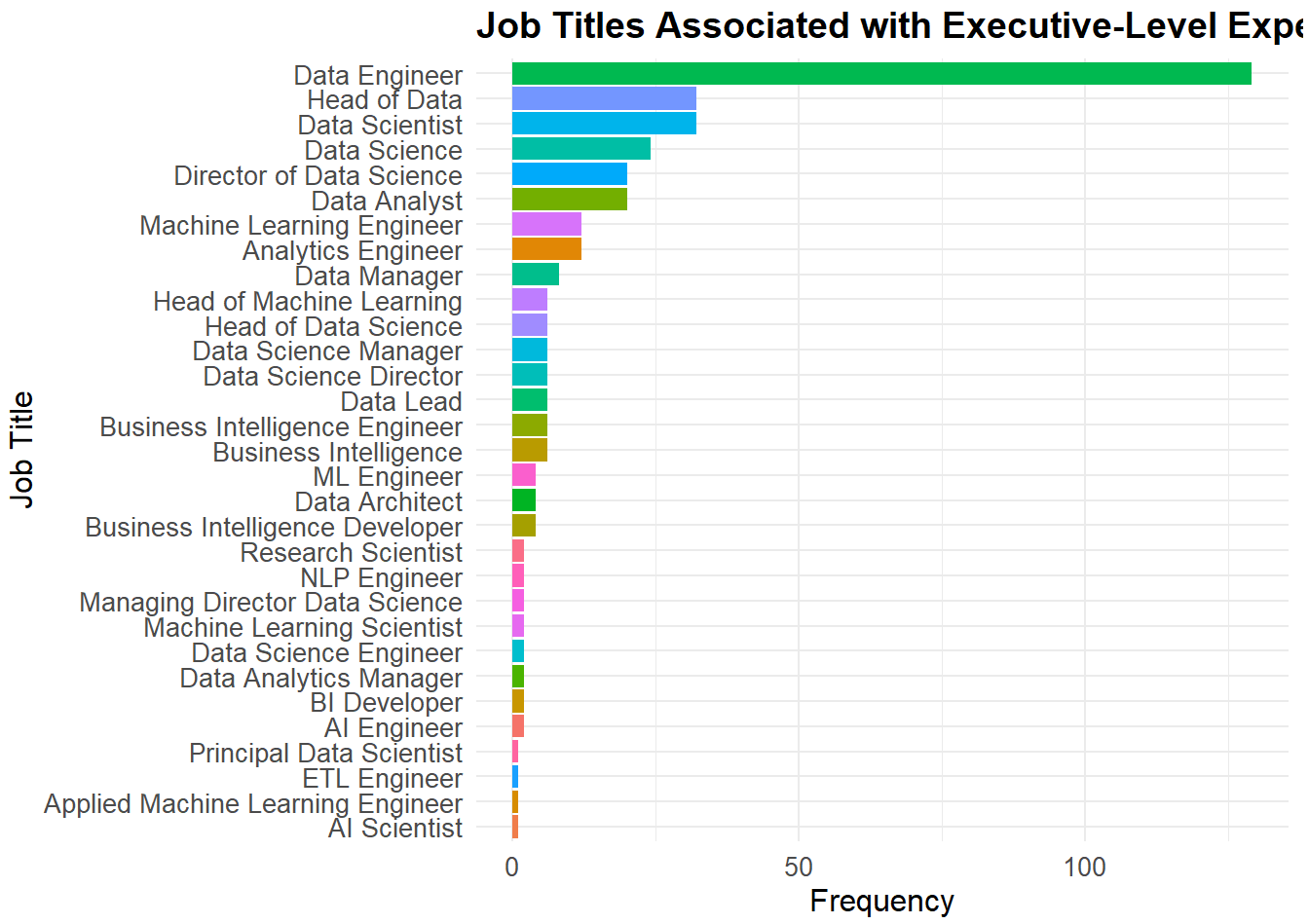
Analysis

```
# Filter data for Executive-level experience
executive_data <- data %>% filter(experience_level == "Executive") %>% count(job_title, name = "Frequency") # Count frequency of each job title
```

Bar chart for job titles with Executive-level experience

```
p <- ggplot(executive_data, aes(x = reorder(job_title, Frequency), y = Frequency, fill = job_tit
le)) +
  geom_bar(stat = "identity", show.legend = FALSE) +
  coord_flip() +
  labs(
    title = "Job Titles Associated with Executive-Level Experience",
    x = "Job Title",
    y = "Frequency",
    fill = "Job Title"
  ) +
  theme_minimal() +
  theme(
    axis.title.x = element_text(size = 12),
    axis.title.y = element_text(size = 12),
    plot.title = element_text(size = 14, face = "bold"),
    axis.text.x = element_text(size = 10),
    axis.text.y = element_text(size = 10)
  )

print(p)
```




```
ggsave("barplot.png", plot = p, width = 8, height = 6, dpi = 300)
```

This chart visualizes the frequency of job titles associated with executive-level experience in a given dataset. The x-axis represents the frequency count, while the y-axis lists various job titles. Prominent roles such as “Data Engineer,” “Head of Data,” and “Data Scientist” are highlighted at the top with the highest frequencies, indicating that these positions are most commonly linked with executive experience. The dominance of technical and leadership-focused roles reflects the demand for specialized expertise combined with management capabilities in data-driven fields.

Lower-frequency job titles, such as “AI Scientist,” “ETL Engineer,” and “Principal Data Scientist,” are also displayed but are less prevalent. These roles may represent niche positions or areas with fewer professionals at the executive level. The chart provides insight into the hierarchical structure of data-related careers and the titles most frequently occupied by individuals at advanced stages of their careers. It highlights the increasing importance of data leadership roles in modern organizations.

What are the trends in salaries over the years?

Introduction: Analyzing trends in salaries over the years reveals how compensation has evolved and provides insight into economic factors, industry demands, and inflation adjustments affecting the job market. Understanding these trends helps identify periods of significant salary growth or stagnation and offers valuable context for career planning and policy decisions.

Approach:

Group the dataset by year (work_year) and calculate the average salary (salary_in_usd) for each year.

Visualize the trends using a line chart, emphasizing yearly changes in average salaries with points marking the averages.

Analysis

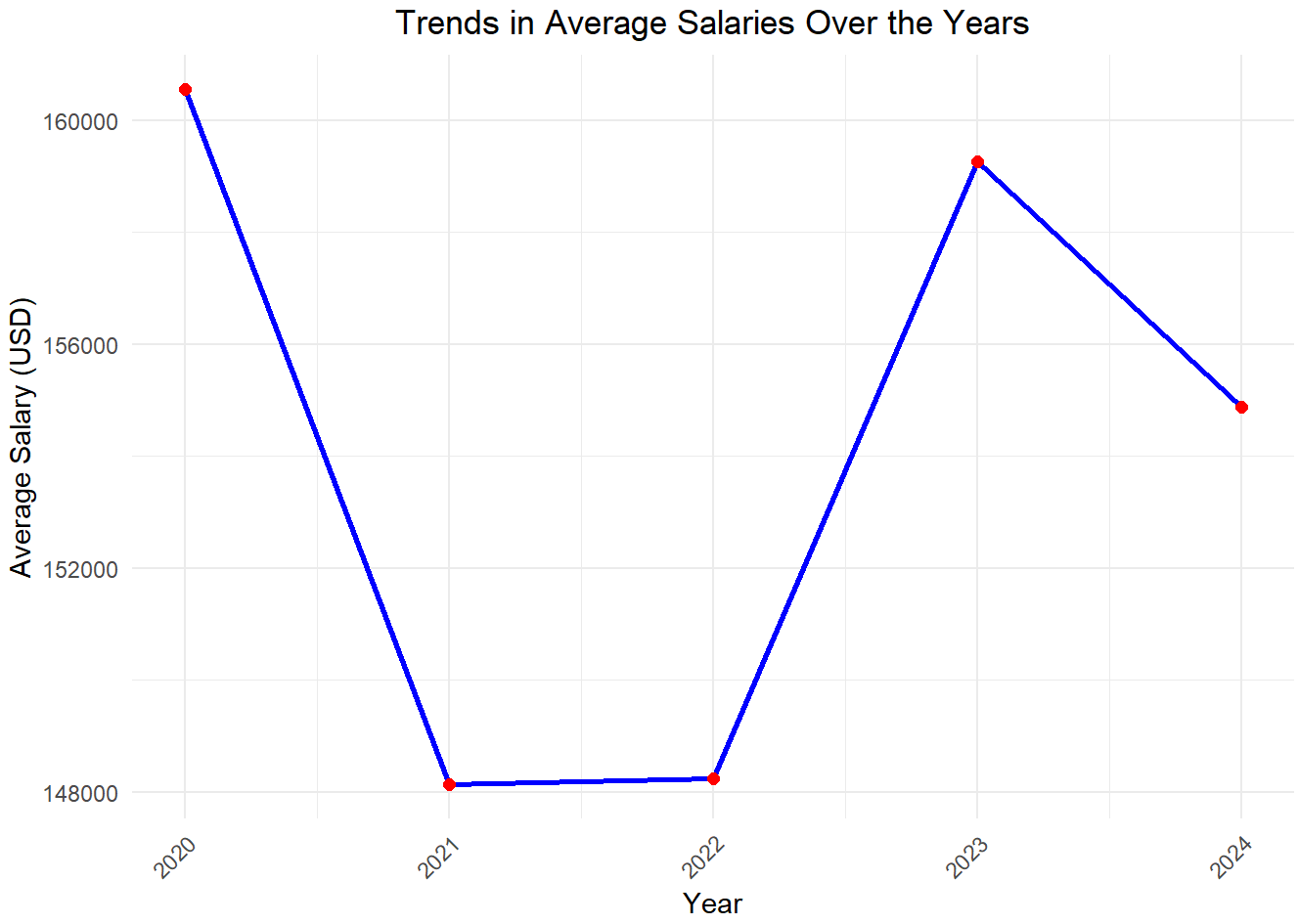
```
# Calculate average salary per year
average_salary <- data %>% group_by(work_year) %>% summarize(avg_salary = mean(salary_in_usd, n
a.rm = TRUE))
```

Create a line chart

```
p <- ggplot(average_salary, aes(x = work_year, y = avg_salary)) +
  geom_line(color = "blue", size = 1) +
  geom_point(color = "red", size = 2) +
  labs(
    title = "Trends in Average Salaries Over the Years",
    x = "Year",
    y = "Average Salary (USD)"
  ) +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5),
    axis.text.x = element_text(angle = 45, hjust = 1)
  )
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## ℹ Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```
print(p)
```




```
ggsave("lineplot.png", plot = p, width = 8, height = 6, dpi = 300)
```

This line chart illustrates the trends in average salaries from 2020 to 2024, showcasing fluctuations over time. The x-axis represents the years, while the y-axis indicates the average salary in USD. Each data point, marked with red dots, highlights the precise salary value for a given year, while the blue line connects these points to reveal the overall trajectory. In 2020, the average salary started at approximately $160,000. However, a sharp decline occurred in 2021, with salaries dropping to around $148,000, marking the lowest point in the observed timeline.

Following this decline, salaries rebounded significantly between 2021 and 2023, peaking again at $160,000 in 2023, reflecting a sharp recovery. However, the trend shows another decline in 2024, although it was less drastic than the drop in 2021. The overall pattern suggests a fluctuation in average salaries, characterized by an initial decline, a recovery period, and a slight downturn in the most recent year. These shifts may reflect broader economic conditions or changes in industry demand over the years.

Approach:

Group the dataset by job_category and company_size, then calculate the average salary (salary_in_usd) for each combination.

Use a multi-line chart to represent salary trends, where each line corresponds to a specific company size and shows how salaries vary by job category.

Analysis

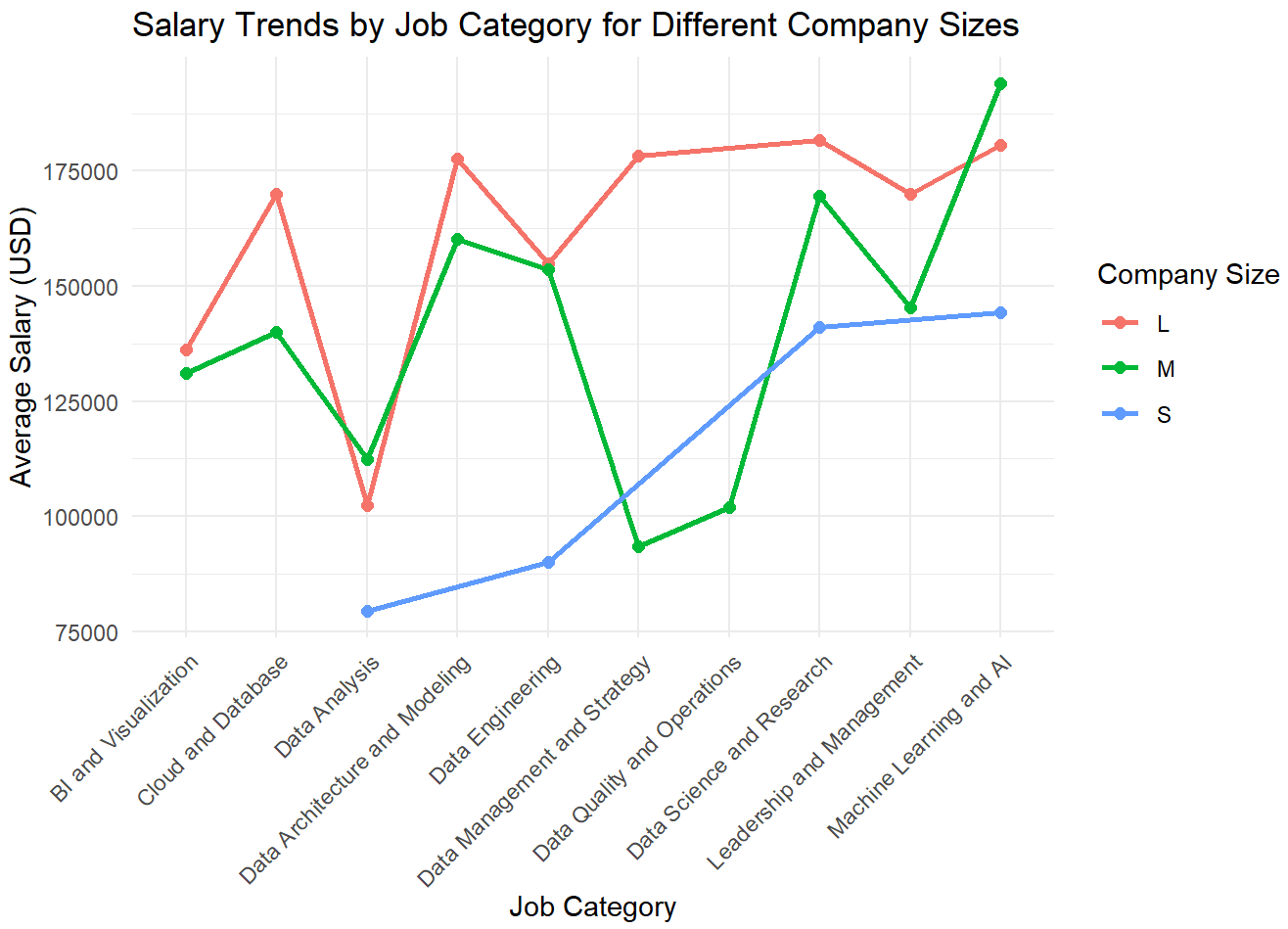
```
# Aggregate data to calculate average salary by job title and company size
salary_trends <- data %>% group_by(job_category, company_size) %>% summarise(avg_salary = mean(s
alary_in_usd, na.rm = TRUE))
```

```
## `summarise()` has grouped output by 'job_category'. You can override using the
## `.groups` argument.
```

Create multi-line chart

```
p <- ggplot(salary_trends, aes(x = job_category, y = avg_salary, color = company_size, group = c
ompany_size)) +
  geom_line(size = 1) +
  geom_point(size = 2) +
  labs(
    title = "Salary Trends by Job Category for Different Company Sizes",
    x = "Job Category",
    y = "Average Salary (USD)",
    color = "Company Size"
  ) +
  theme_minimal() +
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1)
  )

print(p)
```




```
ggsave("multi-lineplot.png", plot = p, width = 8, height = 6, dpi = 300)
```

This multi-line chart illustrates the trends in average salaries across various job categories, segmented by company sizes: large (L), medium (M), and small (S). The x-axis represents the job categories (e.g., BI and Visualization, Data Analysis, Machine Learning and AI), while the y-axis shows the corresponding average salary in USD. Each line represents a company size, with red for large companies, green for medium-sized companies, and blue for small companies. The chart demonstrates salary variations depending on job type and company size.

From the chart, large companies (red line) tend to offer the highest salaries across most job categories, especially for roles such as Data Architecture and Modeling, and Machine Learning and AI, where salaries exceed $175,000. Medium-sized companies (green line) show more fluctuation, with some roles like Data Engineering and Leadership offering competitive salaries but significantly lower for Data Architecture and Modeling. Small companies (blue line) generally offer lower salaries across all categories, though the trend is more consistent, with roles like Machine Learning and AI and Data Science showing noticeable peaks. The data suggests that large organizations invest more in specialized roles, while smaller companies maintain more uniform salary ranges across categories.

How does the work setting correlate with employee residence and salary levels?

Introduction: Understanding the correlation between work settings (e.g., remote, hybrid, in-person), employee residence, and salary levels can reveal how geographic and organizational factors influence compensation. This analysis highlights patterns such as whether remote work leads to higher salaries and how employee location impacts earning potential in different work settings.

Analysis

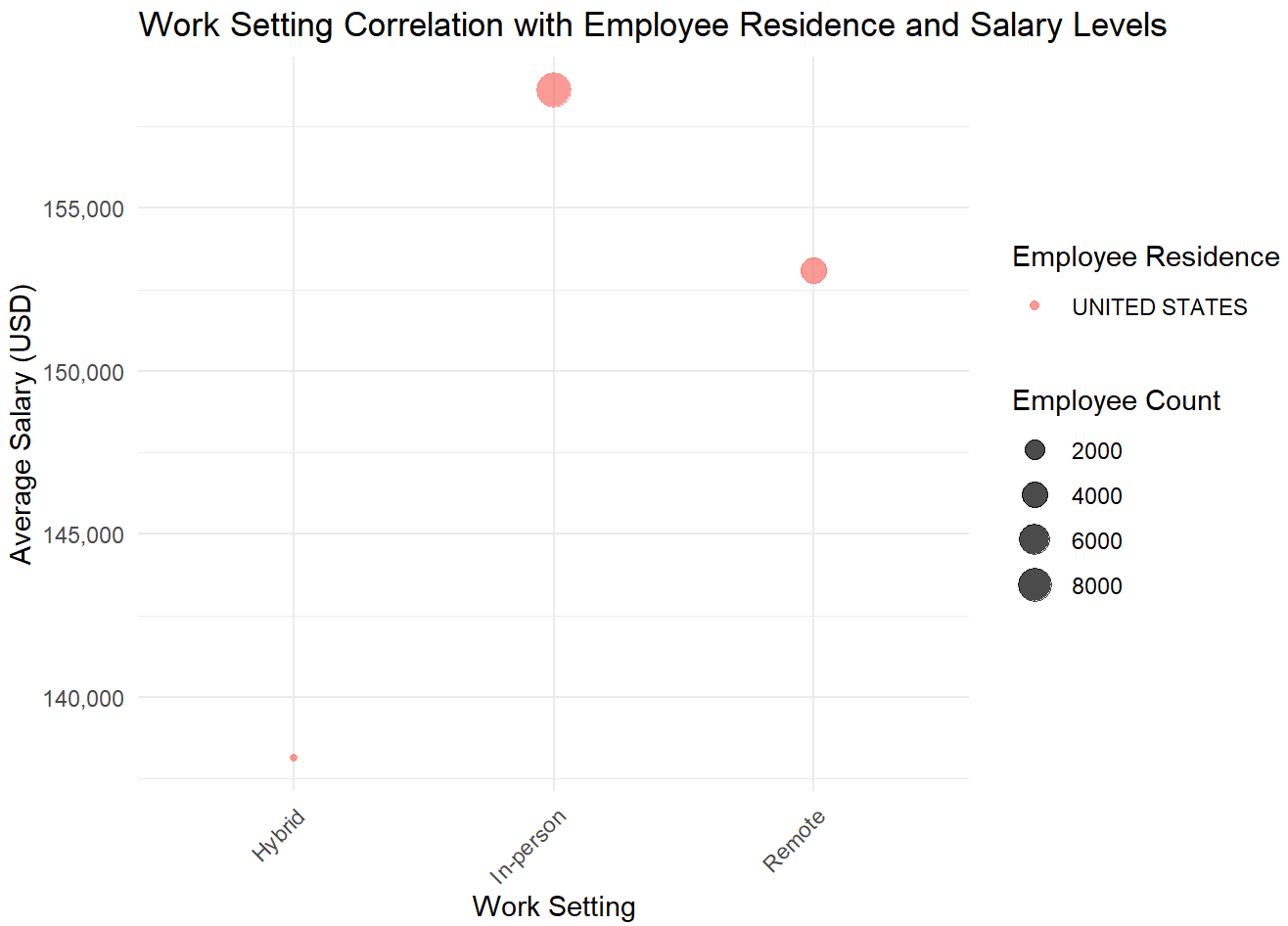
```
# Aggregate data for the chart
bubble_data <- data %>% group_by(work_setting, employee_residence) %>% summarise( avg_salary = m
ean(salary_in_usd, na.rm = TRUE), employee_count = n() )
```

```
## `summarise()` has grouped output by 'work_setting'. You can override using the
## `.groups` argument.
```

Create bubble chart

```
p <- ggplot(bubble_data, aes(x = work_setting, y = avg_salary, size = employee_count, color = em
ployee_residence)) +
  geom_point(alpha = 0.7) +
  labs(
    title = "Work Setting Correlation with Employee Residence and Salary Levels",
    x = "Work Setting",
    y = "Average Salary (USD)",
    size = "Employee Count",
    color = "Employee Residence"
  ) +
  theme_minimal() +
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1)
  ) +
  scale_y_continuous(labels = scales::comma)

print(p)
```




```
ggsave("bubbleplot.png", plot = p, width = 8, height = 6, dpi = 300)
```

This bubble chart examines the relationship between work settings, average salary levels, and employee counts for individuals residing in the United States. The x-axis categorizes work settings into Hybrid, In-person, and Remote, while the y-axis represents average salaries in USD. The size of the bubbles corresponds to the number of employees in each category, and the chart color distinguishes data for employees in the United States.

The chart reveals that the In-person work setting has the largest bubble, indicating the highest number of employees, with an average salary of approximately $155,000, the highest among the three categories. The Hybrid work setting, by contrast, has the smallest bubble, representing fewer employees and the lowest average salary at about $140,000. The Remote work setting falls in between, with a medium-sized bubble and salaries that are higher than Hybrid but lower than In-person. These insights suggest that in-person roles tend to offer higher pay, potentially reflecting employer preferences or the nature of jobs requiring on-site presence.

How have remote work and in-person work settings evolved over the years?

Introduction: The shift between remote and in-person work settings over time reflects broader changes in work culture, technology adoption, and responses to global events. Analyzing the evolution of these trends helps understand the growing prevalence of remote work and its implications for employment practices and employee preferences.

Approach:

Group data by year (work_year) and work setting (work_setting) to calculate the number of employees in each category.

Calculate the proportion of each work setting relative to the total employees for each year.

Use a stacked area chart to illustrate how the proportions of remote and in-person work settings have changed over the years.

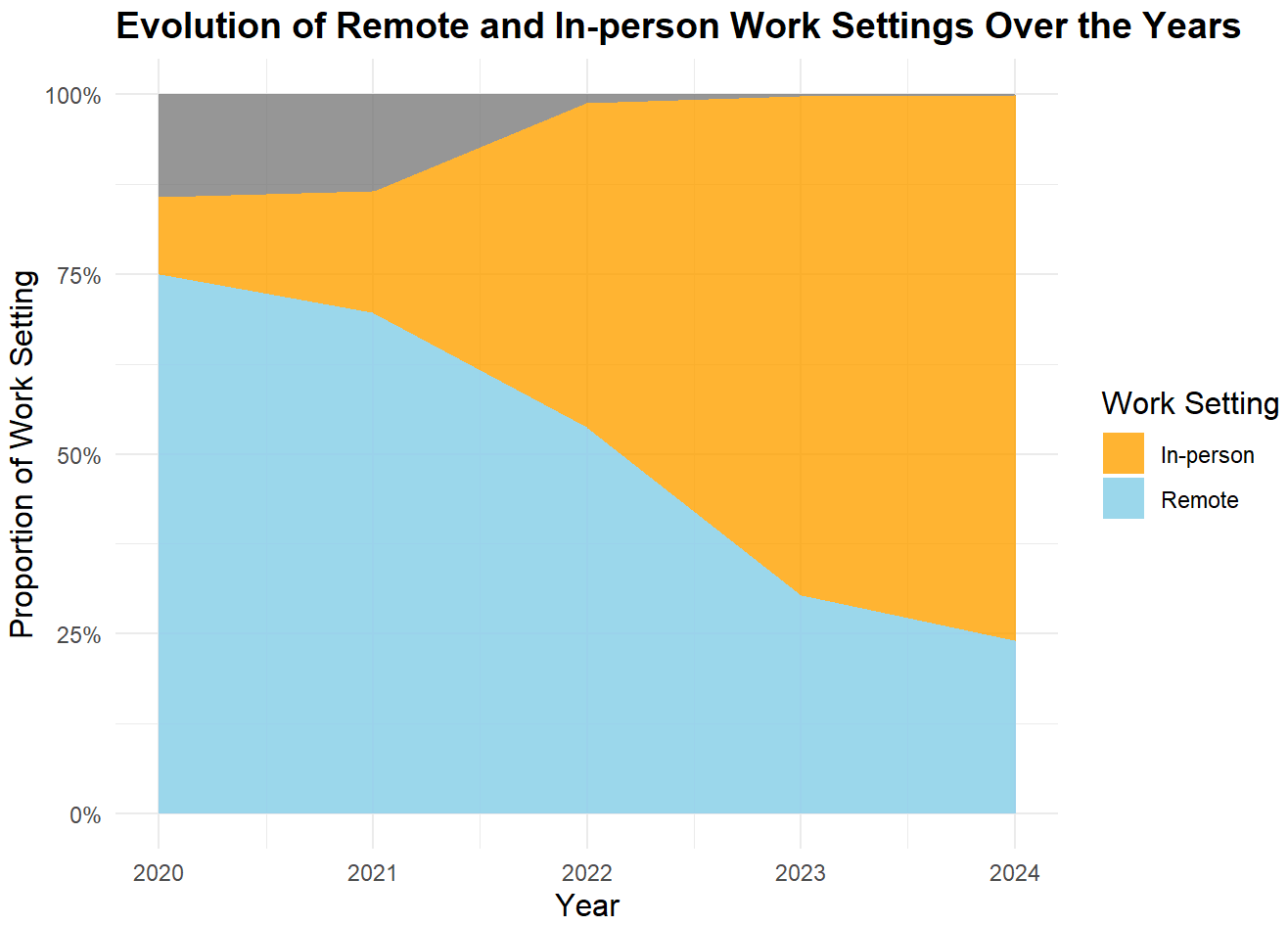
Analysis

```
# Summarize data to count employees per year and work setting
summary_data <- data %>% group_by(work_year, work_setting) %>% summarise(count = n(), .groups =
"drop") %>% group_by(work_year) %>% mutate(proportion = count / sum(count))
```

Stacked area chart

```
p <- ggplot(summary_data, aes(x = work_year, y = proportion, fill = work_setting)) +
  geom_area(alpha = 0.8) +
  scale_fill_manual(values = c("In-person" = "orange", "Remote" = "skyblue")) +
  labs(
    title = "Evolution of Remote and In-person Work Settings Over the Years",
    x = "Year",
    y = "Proportion of Work Setting",
    fill = "Work Setting"
  ) +
  scale_y_continuous(labels = scales::percent) +
  theme_minimal() +
  theme(
    axis.title.x = element_text(size = 12),
    axis.title.y = element_text(size = 12),
    plot.title = element_text(size = 14, face = "bold"),
    legend.title = element_text(size = 12)
  )

print(p)
```




```
ggsave("stackedarea-plot.png", plot = p, width = 8, height = 6, dpi = 300)
```

This graph illustrates the shift in work settings from 2020 to 2024, with a clear focus on the balance between remote and in-person work. Initially, in 2020, in-person work was dominant, but as the pandemic took hold, there was a sharp increase in remote work, peaking around 2021. By 2021, a large proportion of the workforce transitioned to remote settings, reflecting the global response to COVID-19 restrictions. The graph shows that during this period, remote work rapidly overtook in-person work, and the trend remained strong through 2022.

However, starting in 2022, the graph indicates a gradual return to in-person work, as companies began to reopen offices and adapt to new hybrid work models. By 2024, while remote work still holds a significant portion, in-person work has steadily increased, reflecting a shift toward hybrid or fully in-person models in many sectors. The graph captures the evolving work environment in response to both external factors like the pandemic and changing workplace preferences, with remote work declining slightly as in-person work gains ground again.

Discussion and Conclusion

The analysis of job market trends over recent years highlights several key insights that reflect the evolving nature of work environments and compensation structures:

Salaries Across Job Categories and Experience LevelsOur findings reveal significant variations in salaries by job category and experience levels. While higher experience levels, such as executive positions, generally command higher salaries, the presence of outliers shows that mid-level and senior-level positions can occasionally offer salaries comparable to or even exceeding those in executive roles. This trend may result from demand for specialized skills in certain job categories or organizational priorities emphasizing technical expertise over leadership.

Dominant Job Titles in Executive ExperienceCertain job titles are predominantly associated with executive-level experience, indicating a clear hierarchy and specialization within industries. Understanding these roles can guide individuals aiming to reach senior positions while helping organizations identify key leadership profiles.

Yearly Trends in SalariesAverage salaries have shown a noticeable trend over the years, often reflecting economic conditions, demand for specific skills, and global events influencing the labor market. Despite general trends, the analysis highlights variability within specific years, suggesting industry or region-specific factors influencing salaries.

Impact of Company Size on Salary TrendsLarger companies tend to offer higher salaries, especially for certain job categories, likely due to better resources and established pay scales. However, small and medium-sized companies still provide competitive salaries in specific domains, demonstrating that company size is not always the sole determinant of compensation.

Correlations Between Work Setting, Residence, and SalariesThe work setting, whether remote, hybrid, or in-person, correlates with salary levels and employee residence. Remote work, for instance, has opened opportunities for higher salaries independent of geographic constraints, highlighting the growing acceptance and adoption of distributed workforces.

Evolution of Remote and In-Person Work SettingsRemote work has gained prominence in recent years, reflecting shifts in global work culture. This change likely stems from technological advancements and responses to external factors like the COVID-19 pandemic. The increasing proportion of remote work suggests that flexibility has become a critical factor in modern employment dynamics.

Broader Implications:These findings have broader implications for employers and job seekers. Employers can use these insights to design competitive compensation packages and work policies to attract and retain talent. Meanwhile, job seekers can identify roles, industries, and settings aligned with their career aspirations and financial goals.

Limitations:While the dataset is extensive, it primarily focuses on the data science and technology job markets, which may introduce bias in representing trends across other industries. Additionally, the presence of significant outliers in the salary data suggests the need for further refinement or stratification of job categories and experience levels to avoid skewed interpretations. Future research could also explore sector-specific data and more nuanced attributes of remote and hybrid work.

@Conclusion: This project demonstrates the power of data analysis in uncovering patterns and trends in the job market. By understanding how various factors influence salaries and work preferences, stakeholders can make informed decisions that benefit both individuals and organizations. The results also reflect broader trends in the modern workforce, emphasizing the importance of adaptability and innovation in navigating the evolving job landscape. Despite limitations, the findings provide a strong foundation for understanding the dynamics of the job market in a rapidly changing world.

References Kaggle. (2020–2023). Job Market Trends, Salary Data, Job Titles, and Company Sizes. Retrieved from https://www.kaggle.com/datasets/murilozangari/jobs-and-salaries-in-data-field-2024 (https://www.kaggle.com/datasets/murilozangari/jobs-and-salaries-in-data-field-2024)

Monster and Indeed. (2024). Job Availability, Salaries, and Required Skills for Various Positions. Data collected from job listings on Monster and Indeed websites.

Student Survey Responses. (2024). Survey on Career Interests and Job Opportunities. Collected by INFO 526 students for academic research purposes.