Project - Data Science Salaries

Peter Jordan

2025-11-09

Introduction

Our CEO is considering hiring a full-time data scientist to help drive data science across the organization, with the possibility of building a team in the future. The main question is what salary range we should offer to be competitive and attract strong talent.

In this project, I use a global dataset of data science salaries to:

- Describe typical data scientist salaries in USD
- Compare salaries by experience level
- Compare salaries for roles in the United States versus other countries
- Focus on salaries at small companies, which best match our situation
- Recommend a competitive salary range for a full-time data scientist, with an offshore comparison

```
#Load the raw salary data from the CSV file in the project folder
salaries_raw <- read_csv("PeterJordan.module05RProject.csv")
```

```
## New names:
## Rows: 607 Columns: 12
## ── Column specification
## ──────────────────────────────────────────────────────── Delimiter: "," chr
## (7): experience_level, employment_type, job_title, salary_currency, empl... dbl
## (5): ...1, work_year, salary, salary_in_usd, remote_ratio
## ℹ Use `spec()` to retrieve the full column specification for this data. ℹ
## Specify the column types or set `show_col_types = FALSE` to quiet this message.
## • `` -> `...1`
```

```
#Take a quick look at the structure of the data
glimpse(salaries_raw)
```

```
## Rows: 607
## Columns: 12
## $ ...1               <dbl> 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 1…
## $ work_year          <dbl> 2020, 2020, 2020, 2020, 2020, 2020, 2020, 2020, 202…
## $ experience_level   <chr> "MI", "SE", "SE", "MI", "SE", "EN", "SE", "MI", "MI…
## $ employment_type    <chr> "FT", "FT", "FT", "FT", "FT", "FT", "FT", "FT", "FT…
## $ job_title          <chr> "Data Scientist", "Machine Learning Scientist", "Bi…
## $ salary             <dbl> 70000, 260000, 85000, 20000, 150000, 72000, 190000,…
## $ salary_currency    <chr> "EUR", "USD", "GBP", "USD", "USD", "USD", "USD", "H…
## $ salary_in_usd      <dbl> 79833, 260000, 109024, 20000, 150000, 72000, 190000…
## $ employee_residence <chr> "DE", "JP", "GB", "HN", "US", "US", "US", "HU", "US…
## $ remote_ratio       <dbl> 0, 0, 50, 0, 50, 100, 100, 50, 100, 50, 0, 0, 0, 10…
## $ company_location   <chr> "DE", "JP", "GB", "HN", "US", "US", "US", "HU", "US…
## $ company_size       <chr> "L", "S", "M", "S", "L", "L", "S", "L", "L", "S", "…
```

Dataset Metadata

The dataset contains global salary information for data science–related positions.
Each row represents an employee’s reported salary and job information for a specific year.
The metadata for each column are as follows:

- **work_year** – The year the salary was paid.
- **experience_level** – The experience level in the job during the year:
  - EN = Entry-level / Junior
  - MI = Mid-level / Intermediate
  - SE = Senior-level / Expert
  - EX = Executive-level / Director
- **employment_type** – Type of employment:
  - PT = Part-time
  - FT = Full-time
  - CT = Contract
  - FL = Freelance
- **job_title** – The role worked in during the year.
- **salary** – The total gross salary amount paid (in the original currency).
- **salary_currency** – The currency of the salary, expressed as an ISO 4217 code.
- **salary_in_usd** – The salary converted to USD (using yearly FX rates from fxdata.foorilla.com).
- **employee_residence** – The employee’s primary country of residence (ISO 3166 code).
- **remote_ratio** – The amount of work done remotely:
  - 0 = No remote work (<20%)
  - 50 = Partially remote
  - 100 = Fully remote (>80%)
- **company_location** – The country of the employer’s main office or contracting branch (ISO 3166 code).
- **company_size** – The company’s average size during the year:
  - S = Small (<50 employees)
  - M = Medium (50–250 employees)
  - L = Large (>250 employees)

Data Preparation

Before analyzing salaries, I clean and filter the dataset so that it focuses on the roles that are most relevant for our CEO's question. In particular, I:

- Remove the index column created when the data were exported
- Filter to full-time roles
- Keep all job titles, since the dataset already represents data science–related positions
- Create a simple indicator for US vs Non-US company locations
- Treat the experience level as an ordered factor (EN, MI, SE, EX) for easier comparison across levels

```
#Clean and filter the data

salaries_clean <- salaries_raw %>%
  # Remove index column
  select(-...1) %>%
  # Keep full-time roles only
  filter(employment_type == "FT") %>%
  # Recode experience level and create US vs Non-US flag
  mutate(
    experience_level = factor(
      experience_level,
      levels = c("EN", "MI", "SE", "EX"),
      ordered = TRUE
    ),
    us_vs_nonus = if_else(company_location == "US", "US", "Non-US")
  )

#Check the cleaned data
glimpse(salaries_clean)
```

```
## Rows: 588
## Columns: 12
## $ work_year          <dbl> 2020, 2020, 2020, 2020, 2020, 2020, 2020, 2020, 202…
## $ experience_level   <ord> MI, SE, SE, MI, SE, EN, SE, MI, MI, SE, EN, MI, EN,…
## $ employment_type    <chr> "FT", "FT", "FT", "FT", "FT", "FT", "FT", "FT", "FT…
## $ job_title          <chr> "Data Scientist", "Machine Learning Scientist", "Bi…
## $ salary             <dbl> 70000, 260000, 85000, 20000, 150000, 72000, 190000,…
## $ salary_currency    <chr> "EUR", "USD", "GBP", "USD", "USD", "USD", "USD", "H…
## $ salary_in_usd      <dbl> 79833, 260000, 109024, 20000, 150000, 72000, 190000…
## $ employee_residence <chr> "DE", "JP", "GB", "HN", "US", "US", "US", "HU", "US…
## $ remote_ratio       <dbl> 0, 0, 50, 0, 50, 100, 100, 50, 100, 50, 0, 0, 0, 10…
## $ company_location   <chr> "DE", "JP", "GB", "HN", "US", "US", "US", "HU", "US…
## $ company_size       <chr> "L", "S", "M", "S", "L", "L", "S", "L", "L", "S", "…
## $ us_vs_nonus        <chr> "Non-US", "Non-US", "Non-US", "Non-US", "US", "US",…
```

Salary summary function

To keep the code organized and avoid repeating the same summary logic, I define a small helper function that computes basic summary statistics for salary_in_usd. I will reuse this function throughout the analysis for different groups (overall, by experience level, by location, etc.).

```r
#Helper function to summarize salary_in_usd in a consistent way
summarise_salaries <- function(data) {
  data %>%
    summarise(
      n = n(),
      min_salary   = min(salary_in_usd, na.rm = TRUE),
      q1_salary    = quantile(salary_in_usd, 0.25, na.rm = TRUE),
      median_salary= median(salary_in_usd, na.rm = TRUE),
      mean_salary  = mean(salary_in_usd, na.rm = TRUE),
      q3_salary    = quantile(salary_in_usd, 0.75, na.rm = TRUE),
      max_salary   = max(salary_in_usd, na.rm = TRUE)
    )
}
```

Overall Salary Overview

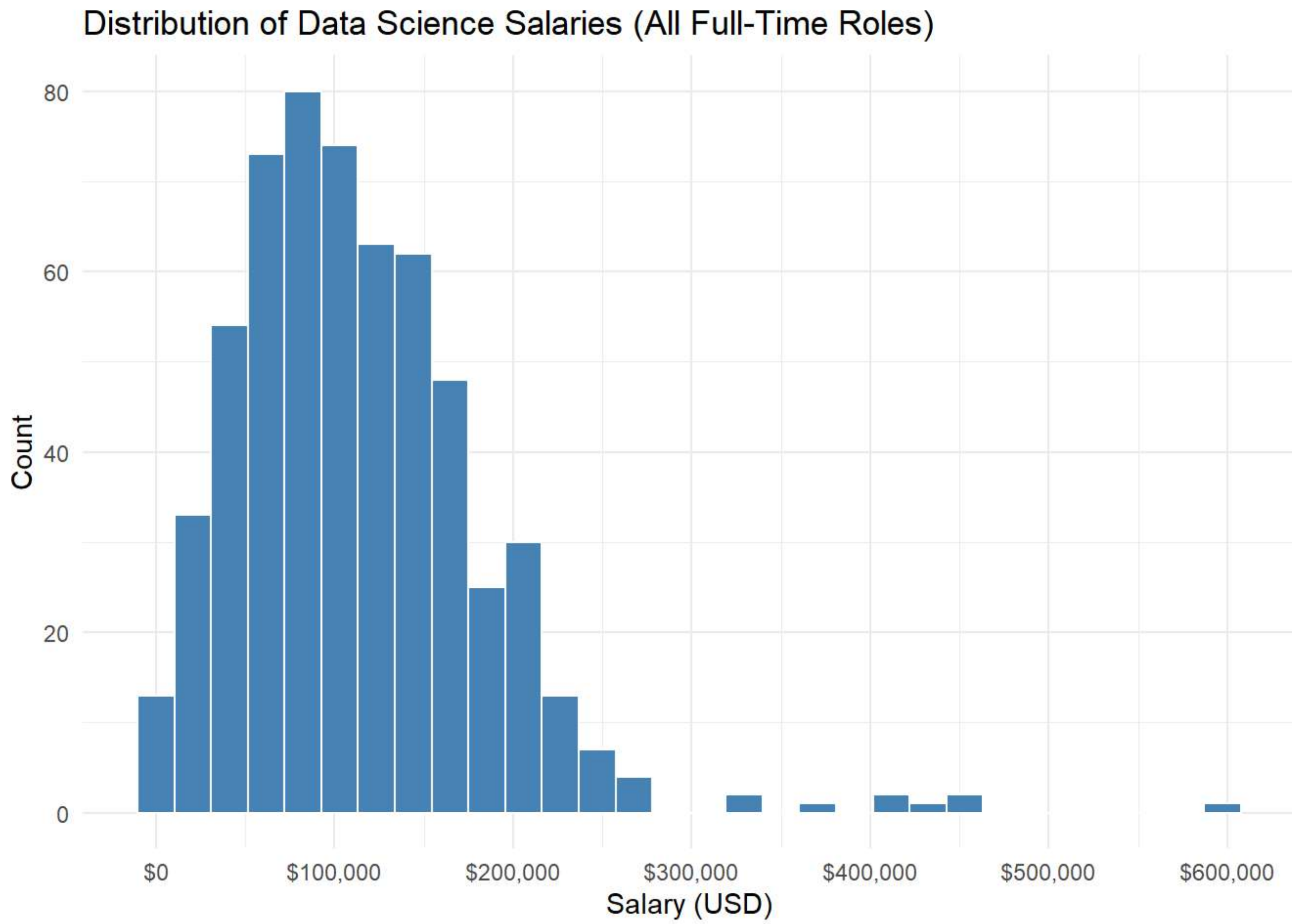
Before comparing by experience or location, I first summarize the overall distribution of salaries for all full-time data-related roles in the dataset.

```r
# Summarize the overall salary distribution in USD
overall_salary_summary <- summarise_salaries(salaries_clean)
overall_salary_summary
```

```
## # A tibble: 1 × 7
##       n min_salary q1_salary median_salary mean_salary q3_salary max_salary
##   <int>      <dbl>     <dbl>         <dbl>       <dbl>     <dbl>      <dbl>
## 1   588       2859    64962.       104196.     113468.    150000     600000
```

The overall median salary for full-time data science roles is approximately $104,000, with an average around $113,000.
The middle 50% of salaries fall roughly between $65,000 and $150,000, showing a wide range across experience levels and countries.
The minimum and maximum values indicate that the dataset includes both entry-level and high-seniority positions.
This overview provides useful context before comparing how experience and location affect pay.

```r
#Visualize the distribution of salaries
ggplot(salaries_clean, aes(x = salary_in_usd)) +
  geom_histogram(bins = 30, fill = "steelblue", color = "white") +
  scale_x_continuous(
    labels = scales::dollar_format(),
    breaks = seq(0, 600000, by = 100000)
  ) +
  labs(
    title = "Distribution of Data Science Salaries (All Full-Time Roles)",
    x = "Salary (USD)",
    y = "Count"
  ) +
  theme_minimal()
```



#

Experience-Level Analysis

Next, I examine how salaries differ by experience level.
This helps identify how much more companies pay for mid-level, senior, and executive roles compared to entry-level positions.

```
#Summarize salaries by experience level
salary_by_experience <- salaries_clean %>%
  group_by(experience_level) %>%
  summarise_salaries()

salary_by_experience %>%
  mutate(across(where(is.numeric), ~ scales::comma(round(., 0))))
```

```
## # A tibble: 4 × 8
##   experience_level n     min_salary q1_salary median_salary mean_salary
##   <ord>            <chr> <chr>      <chr>     <chr>         <chr>
## 1 EN               79    4,000      33,536    59,102        64,457
## 2 MI               206   2,859      49,461    77,161        88,403
## 3 SE               278   18,907     100,000   136,300       139,021
## 4 EX               25    69,741     130,000   167,875       190,728
## # ℹ 2 more variables: q3_salary <chr>, max_salary <chr>
```

```
ggplot(salaries_clean, aes(x = experience_level, y = salary_in_usd)) +
  geom_boxplot(fill = "skyblue") +
  scale_y_continuous(
    labels = scales::dollar_format(),
    breaks = seq(0, 600000, by = 100000)
  ) +
  labs(
    title = "Data Science Salaries by Experience Level",
    x = "Experience Level",
    y = "Salary (USD)"
  ) +
  theme_minimal()
```

The boxplot shows a clear upward trend in salaries as experience increases.
Entry-level (EN) employees typically earn below $100,000, while mid-level (MI) and senior-level (SE) professionals cluster between $100,000 and $175,000.

Executive-level (EX) roles show the highest range, often exceeding $200,000 and reaching up to $600,000 for the most senior positions.
This pattern demonstrates that experience has a strong positive relationship with salary in data science roles.

U.S. vs. Non-U.S. Salary Comparison

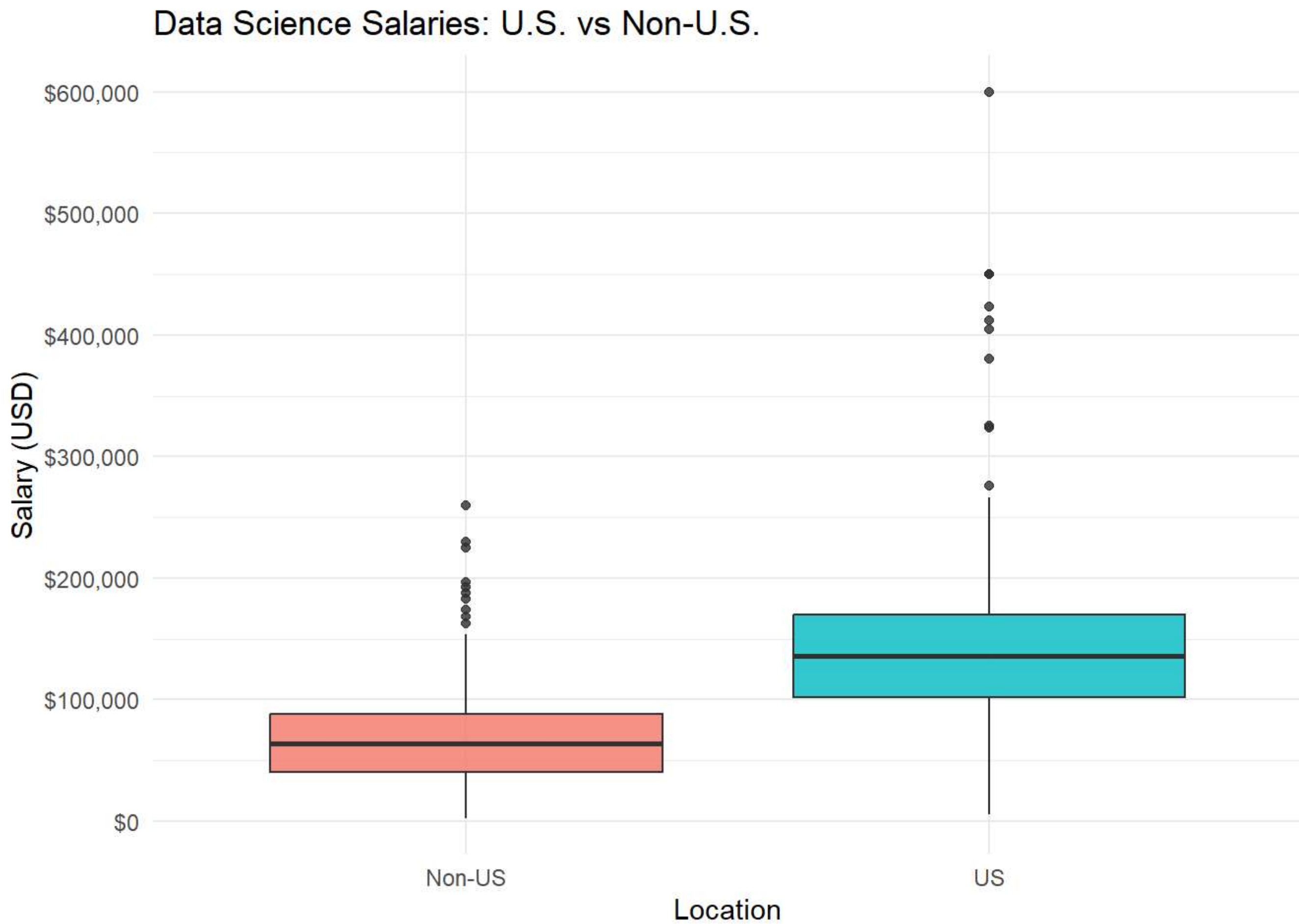
The CEO asked how data science salaries differ between employees based in the United States and those located offshore.
This section compares full-time salaries (in USD) for both groups to highlight the pay gap and help guide hiring decisions.

```
#Summarize salaries for U.S. vs Non-U.S.
salary_us_vs_nonus <- salaries_clean %>%
  group_by(us_vs_nonus) %>%
  summarise_salaries()

#Display the summary nicely formatted
salary_us_vs_nonus %>%
  mutate(across(where(is.numeric), ~ scales::comma(round(., 0))))
```

```
## # A tibble: 2 × 8
##   us_vs_nonus n     min_salary q1_salary median_salary mean_salary q3_salary
##   <chr>       <chr> <chr>      <chr>     <chr>         <chr>       <chr>
## 1 Non-US      242   2,859      40,262    63,760        68,903      88,474
## 2 US          346   5,679      102,100   136,300       144,638     170,000
## # ℹ 1 more variable: max_salary <chr>
```

```
ggplot(salaries_clean, aes(x = us_vs_nonus, y = salary_in_usd, fill = us_vs_nonus)) +
  geom_boxplot(alpha = 0.8) +
  scale_y_continuous(
    labels = scales::dollar_format(),
    breaks = seq(0, 600000, by = 100000)
  ) +
  labs(
    title = "Data Science Salaries: U.S. vs Non-U.S.",
    x = "Location",
    y = "Salary (USD)",
    fill = "Region"
  ) +
  theme_minimal() +
  theme(legend.position = "none")
```



The U.S. and Non-U.S. comparison shows that domestic salaries generally exceed $100,000, while most Non-U.S. salaries fall below this level.
This aligns with expectations given higher cost of living and competition for technical roles in the United States.
The chart also reinforces that U.S.-based professionals command a premium for experience and proximity, whereas offshore hiring offers significant cost savings but may trade off convenience and leadership potential.

Small-Company Salary Analysis and Recommendation

Because the company is still small but expanding, it’s important to understand what similar-sized organizations pay for data science roles.
This section filters the dataset to small companies (less than 50 employees) and compares U.S. and Non-U.S. salaries to recommend a competitive range.

```
#Filter to small companies only
small_company_salaries <- salaries_clean %>%
  filter(company_size == "S")

#Summarize small-company salaries by region
small_company_summary <- small_company_salaries %>%
  group_by(us_vs_nonus) %>%
  summarise_salaries()

#Display clean summary
small_company_summary %>%
  mutate(across(where(is.numeric), ~ scales::comma(round(., 0))))
```

```
## # A tibble: 2 × 8
##   us_vs_nonus n     min_salary q1_salary median_salary mean_salary q3_salary
##   <chr>       <chr> <chr>      <chr>     <chr>         <chr>       <chr>
## 1 Non-US      49    2,859      25,532    62,726        63,991      77,364
## 2 US          28    5,679      59,500    90,000        98,346      120,000
## # ℹ 1 more variable: max_salary <chr>
```

```
ggplot(small_company_salaries, aes(x = us_vs_nonus, y = salary_in_usd, fill = us_vs_nonus)) +
  geom_boxplot(alpha = 0.8) +
  scale_y_continuous(
    labels = scales::dollar_format(),
    breaks = seq(0, 600000, by = 100000)
  ) +
  labs(
    title = "Small-Company Data Science Salaries: U.S. vs Non-U.S.",
    x = "Location",
    y = "Salary (USD)",
    fill = "Region"
  ) +
  theme_minimal() +
  theme(legend.position = "none")
```

For small companies, U.S.-based data scientists earn a median salary close to $110,000, with the top quartile approaching $130,000–$150,000.
Non-U.S. small-company salaries are considerably lower, averaging around $70,000–$90,000.
This gap highlights that while offshore hiring offers significant cost savings, hiring within the U.S. provides access to top-tier candidates and better alignment as the company scales.
Based on these insights, a **competitive and sustainable range** for the company would be:

- **U.S. hire:** $110,000 – $140,000
- **Offshore hire:** $70,000 – $90,000

Conclusion

This analysis demonstrates that experience level, company size, and geography all have a strong impact on data science salaries.
For a small but growing organization, offering a range of **$110K–$140K** for a U.S.-based data scientist or **$70K–$90K** for an offshore professional will keep the company competitive in attracting top talent while managing budget efficiently.
These recommendations provide a foundation for sustainable growth as the company expands its data capabilities.