Project\_1

Beija

2025-04-19

Due to a shared understanding that salaries for data scientists vary significantly across the globe, as well as additional factors such as the Great Recession and market competitiveness, what is the salary range necessary to attract top talent for positions within the United States?

What are the cost differences between domestic and offshore hires?

What salary should a growing company offer in order to attract top data scientist talent, whether based in the United States or offshore, within the context of today’s competitive market?

What is the competitive salary range for a full-time data scientist in the United States compared to other global regions?

Additionally, it would be beneficial to specify the salary ranges for both entry-level and senior-level positions, as these distinctions significantly influence hiring decisions.

library(readr)  
library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

salary\_data <- read\_csv("r project data-1.csv")

## New names:  
## • `` -> `...1`

## 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.

str(salary\_data)

## spc\_tbl\_ [607 × 12] (S3: spec\_tbl\_df/tbl\_df/tbl/data.frame)  
## $ ...1 : num [1:607] 0 1 2 3 4 5 6 7 8 9 ...  
## $ work\_year : num [1:607] 2020 2020 2020 2020 2020 2020 2020 2020 2020 2020 ...  
## $ experience\_level : chr [1:607] "MI" "SE" "SE" "MI" ...  
## $ employment\_type : chr [1:607] "FT" "FT" "FT" "FT" ...  
## $ job\_title : chr [1:607] "Data Scientist" "Machine Learning Scientist" "Big Data Engineer" "Product Data Analyst" ...  
## $ salary : num [1:607] 70000 260000 85000 20000 150000 72000 190000 11000000 135000 125000 ...  
## $ salary\_currency : chr [1:607] "EUR" "USD" "GBP" "USD" ...  
## $ salary\_in\_usd : num [1:607] 79833 260000 109024 20000 150000 ...  
## $ employee\_residence: chr [1:607] "DE" "JP" "GB" "HN" ...  
## $ remote\_ratio : num [1:607] 0 0 50 0 50 100 100 50 100 50 ...  
## $ company\_location : chr [1:607] "DE" "JP" "GB" "HN" ...  
## $ company\_size : chr [1:607] "L" "S" "M" "S" ...  
## - attr(\*, "spec")=  
## .. cols(  
## .. ...1 = col\_double(),  
## .. work\_year = col\_double(),  
## .. experience\_level = col\_character(),  
## .. employment\_type = col\_character(),  
## .. job\_title = col\_character(),  
## .. salary = col\_double(),  
## .. salary\_currency = col\_character(),  
## .. salary\_in\_usd = col\_double(),  
## .. employee\_residence = col\_character(),  
## .. remote\_ratio = col\_double(),  
## .. company\_location = col\_character(),  
## .. company\_size = col\_character()  
## .. )  
## - attr(\*, "problems")=<externalptr>

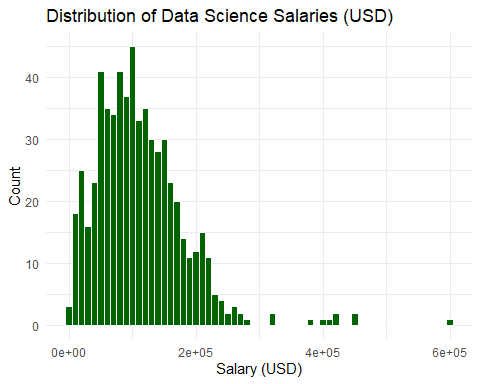
# Convert to factors  
salary\_data$experience\_level <- as.factor(salary\_data$experience\_level)  
salary\_data$employment\_type <- as.factor(salary\_data$employment\_type)  
salary\_data$job\_title <- as.factor(salary\_data$job\_title)  
salary\_data$salary\_currency <- as.factor(salary\_data$salary\_currency)  
salary\_data$employee\_residence <- as.factor(salary\_data$employee\_residence)  
salary\_data$company\_location <- as.factor(salary\_data$company\_location)  
salary\_data$company\_size <- as.factor(salary\_data$company\_size)

Data Analysis

library (ggplot2)  
summary(salary\_data)

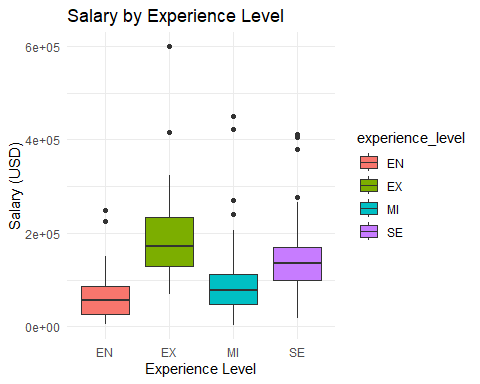
## ...1 work\_year experience\_level employment\_type  
## Min. : 0.0 Min. :2020 EN: 88 CT: 5   
## 1st Qu.:151.5 1st Qu.:2021 EX: 26 FL: 4   
## Median :303.0 Median :2022 MI:213 FT:588   
## Mean :303.0 Mean :2021 SE:280 PT: 10   
## 3rd Qu.:454.5 3rd Qu.:2022   
## Max. :606.0 Max. :2022   
##   
## job\_title salary salary\_currency  
## Data Scientist :143 Min. : 4000 USD :398   
## Data Engineer :132 1st Qu.: 70000 EUR : 95   
## Data Analyst : 97 Median : 115000 GBP : 44   
## Machine Learning Engineer: 41 Mean : 324000 INR : 27   
## Research Scientist : 16 3rd Qu.: 165000 CAD : 18   
## Data Science Manager : 12 Max. :30400000 JPY : 3   
## (Other) :166 (Other): 22   
## salary\_in\_usd employee\_residence remote\_ratio company\_location  
## Min. : 2859 US :332 Min. : 0.00 US :355   
## 1st Qu.: 62726 GB : 44 1st Qu.: 50.00 GB : 47   
## Median :101570 IN : 30 Median :100.00 CA : 30   
## Mean :112298 CA : 29 Mean : 70.92 DE : 28   
## 3rd Qu.:150000 DE : 25 3rd Qu.:100.00 IN : 24   
## Max. :600000 FR : 18 Max. :100.00 FR : 15   
## (Other):129 (Other):108   
## company\_size  
## L:198   
## M:326   
## S: 83   
##   
##   
##   
##

ggplot(salary\_data, aes(x = salary\_in\_usd)) +  
 geom\_histogram(binwidth = 10000, fill = "darkgreen", color = "white") +  
 labs(title = "Distribution of Data Science Salaries (USD)", x = "Salary (USD)", y = "Count") +  
 theme\_minimal()



The graph illustrating the Distribution of Data Science Salaries (USD) indicates that the highest concentration of salaries is observed between $60,000 and $150,000 USD. The modal salary range appears to fall within $100,000 to $130,000, which is likely indicative of the typical salary for data scientists on a global scale. Furthermore, there exists a small subset of individuals who earn significantly high salaries, exceeding $300,000, resulting in a long tail on the right side of the distribution. In order to offer a competitive salary, it is advisable for the CEO to consider a starting range of approximately $100,000 to $150,000 USD, contingent upon the specific location and level of experience of the candidates.

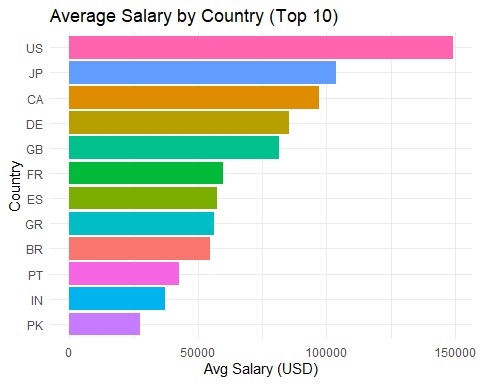
ggplot(salary\_data, aes(x = experience\_level, y = salary\_in\_usd, fill = experience\_level)) +  
 geom\_boxplot() +  
 labs(title = "Salary by Experience Level", x = "Experience Level", y = "Salary (USD)") +  
 theme\_minimal()

 The salary data classified by experience level indicates that Executives (EX) command the highest compensation, demonstrating a broad range of salaries, with several significant high outliers reaching as much as $600,000 to $800,000. Furthermore, there is a discernible salary progression associated with increasing levels of experience:

* **Entry-Level (EN)**: The median salary is approximately $50,000 to $80,000.
* **Mid-Level (MI)**: The median salary is around $100,000.
* **Senior-Level (SE)**: The median salary is approximately $150,000.
* **Executive (EX)**: The median salary typically ranges from $200,000 to $250,000.

For candidates anticipated to lead a team or assume the position of head of data, it would be prudent to offer a compensation package in the range of $180,000 to $250,000 in order to remain competitive, particularly for positions based in the United States.

# Top 10 employee residence locations by count  
top\_countries <- salary\_data %>%  
 count(employee\_residence, sort = TRUE) %>%  
 top\_n(10, n) %>%  
 pull(employee\_residence)  
  
# Filtered dataset  
filtered\_data <- salary\_data %>%   
 filter(employee\_residence %in% top\_countries)  
  
# Bar plot of average salary by country  
filtered\_data %>%  
 group\_by(employee\_residence) %>%  
 summarise(avg\_salary = mean(salary\_in\_usd)) %>%  
 ggplot(aes(x = reorder(employee\_residence, avg\_salary), y = avg\_salary, fill = employee\_residence)) +  
 geom\_col(show.legend = FALSE) +  
 coord\_flip() +  
 labs(title = "Average Salary by Country (Top 10)", x = "Country", y = "Avg Salary (USD)") +  
 theme\_minimal()



The average salary by county in the United States (U.S.) leads globally, with the highest averages exceeding $140,000 USD. This figure underscores the premium associated with U.S.-based talent.

Japan (JP) and Canada (CA) also demonstrate high average salaries, generally exceeding $100,000 USD. In comparison, Germany (DE), the United Kingdom (GB), France (FR), and Spain (ES) fall into a middle salary range.

Conversely, Brazil (BR), Portugal (PT), India (IN), and Pakistan (PK) exhibit considerably lower average salaries. These nations present compelling options for cost-effective offshore hiring; however, variations in experience levels and market maturity should be considered.

Engaging talent within the U.S. typically necessitates a premium investment, often starting at $140,000 USD for experienced professionals. To manage costs while preserving quality, organizations may wish to explore nearshore or offshore alternatives in Europe or Asia, where salaries generally range from $30,000 to $90,000 USD.

Establishing a hybrid team that combines U.S. leadership with offshore support can effectively achieve a balance between quality and cost efficiency.

print(head(salary\_data))

## # A tibble: 6 × 12  
## ...1 work\_year experience\_level employment\_type job\_title salary  
## <dbl> <dbl> <fct> <fct> <fct> <dbl>  
## 1 0 2020 MI FT Data Scientist 70000  
## 2 1 2020 SE FT Machine Learning Scie… 260000  
## 3 2 2020 SE FT Big Data Engineer 85000  
## 4 3 2020 MI FT Product Data Analyst 20000  
## 5 4 2020 SE FT Machine Learning Engi… 150000  
## 6 5 2020 EN FT Data Analyst 72000  
## # ℹ 6 more variables: salary\_currency <fct>, salary\_in\_usd <dbl>,  
## # employee\_residence <fct>, remote\_ratio <dbl>, company\_location <fct>,  
## # company\_size <fct>