bruder_module05_project01

Geoffrey Bruder

2024-10-14

Contents

Global Data Science Salaries: Navigating Wide Ranges and Remote Work Considerations to Attract Top Talent in a Competitive Market

A visual analysis of Salary Distribution: U.S.A. versus Offshore Average Salary by Experience Level: U.S.A. versus Offshore Salary versus Remote Work Ratio

Prepared by: Geoffrey Bruder

```r
# Load necessary libraries
library(tidyr)
library(ggplot2)
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
```

```r
library(readr)
library(knitr)
```

```r
# Read the data
data_path <- "/Users/Main/Documents/DSE5002/r project data.csv"
data <- read_csv(data_path)
```

```
## 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`
```

```r
# Filter data for full-time positions
full_time_data <- data %>% filter(employment_type == "FT")

# Summarize the full_time_data dataframe
summary(full_time_data)
```

```
##       ...1         work_year    experience_level   employment_type
##  Min.   :  0.0   Min.   :2020   Length:588         Length:588
##  1st Qu.:155.8   1st Qu.:2021   Class :character   Class :character
##  Median :308.5   Median :2022   Mode  :character   Mode  :character
##  Mean   :306.0   Mean   :2021
##  3rd Qu.:455.2   3rd Qu.:2022
##  Max.   :606.0   Max.   :2022
##   job_title             salary         salary_currency    salary_in_usd
##  Length:588         Min.   :    4000   Length:588         Min.   :  2859
##  Class :character   1st Qu.:   70000   Class :character   1st Qu.: 64962
##  Mode  :character   Median :  115250   Mode  :character   Median :104196
##                     Mean   :  331125                      Mean   :113468
##                     3rd Qu.:  165000                      3rd Qu.:150000
##                     Max.   :30400000                      Max.   :600000
##  employee_residence  remote_ratio    company_location   company_size
##  Length:588         Min.   :  0.00   Length:588         Length:588
##  Class :character   1st Qu.: 50.00   Class :character   Class :character
##  Mode  :character   Median :100.00   Mode  :character   Mode  :character
##                     Mean   : 70.75
##                     3rd Qu.:100.00
##                     Max.   :100.00
```

```r
# Display the first few rows of the full_time_data dataframe
head(full_time_data)
```

```
## # A tibble: 6 x 12
##    ...1 work_year experience_level employment_type job_title              salary
##   <dbl>     <dbl> <chr>            <chr>           <chr>                   <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 <chr>, salary_in_usd <dbl>,
## #   employee_residence <chr>, remote_ratio <dbl>, company_location <chr>,
## #   company_size <chr>
```

```r
# List the column names of the full_time_data dataframe
colnames(full_time_data)
```

```
##  [1] "...1"               "work_year"          "experience_level"
##  [4] "employment_type"    "job_title"          "salary"
##  [7] "salary_currency"    "salary_in_usd"      "employee_residence"
## [10] "remote_ratio"       "company_location"   "company_size"
```

```r
# Convert the column names of full_time_data dataframe to a list
colnames_list <- as.list(colnames(full_time_data))
```

```r
# Summary statistics for U.S. vs. offshore salaries
us_salaries <- full_time_data %>% filter(employee_residence == "US")
offshore_salaries <- full_time_data %>% filter(employee_residence != "US")

# Summary stats
us_summary <- us_salaries %>% summarize(
  avg_salary = mean(salary_in_usd, na.rm = TRUE),
  median_salary = median(salary_in_usd, na.rm = TRUE),
  sd_salary = sd(salary_in_usd, na.rm = TRUE)
)

offshore_summary <- offshore_salaries %>% summarize(
  avg_salary = mean(salary_in_usd, na.rm = TRUE),
  median_salary = median(salary_in_usd, na.rm = TRUE),
  sd_salary = sd(salary_in_usd, na.rm = TRUE)
)
```

```
# Reformat the avg_salary, median_salary, and sd_salary columns
us_summary <- us_summary %>%
  mutate(
    avg_salary = paste0("$", formatC(avg_salary, format = "f", digits = 2)),
    median_salary = paste0("$", formatC(median_salary, format = "f", digits = 2)),
    sd_salary = paste0("$", formatC(sd_salary, format = "f", digits = 2))
  )

# View the reformatted dataframe
print(us_summary)
```

```
## # A tibble: 1 x 3
##   avg_salary median_salary sd_salary
##   <chr>      <chr>         <chr>
## 1 $148297.09 $138475.00    $66655.66
```

```
# Reformat the avg_salary, median_salary, and sd_salary columns
offshore_summary <- offshore_summary %>%
  mutate(
    avg_salary = paste0("$", formatC(avg_salary, format = "f", digits = 2)),
    median_salary = paste0("$", formatC(median_salary, format = "f", digits = 2)),
    sd_salary = paste0("$", formatC(sd_salary, format = "f", digits = 2))
  )

# View the reformatted dataframe
print(offshore_summary)
```

```
## # A tibble: 1 x 3
##   avg_salary median_salary sd_salary
##   <chr>      <chr>         <chr>
## 1 $69529.92  $63760.50     $43083.56
```

Key Salary Metrics for Analyzing Compensation Data

```
cat("Based on the analysis:
    For U.S.-based data scientists:
    average salary:", us_summary$avg_salary,
    "median salary:", us_summary$median_salary,
    "standard deviation", us_summary$sd_salary,

    "For Offshore-based data scientists:,
    average salary:", offshore_summary$avg_salary,
    "median salary:", offshore_summary$median_salary,
    "standard deviation:", offshore_summary$sd_salary
)
```

```
## Based on the analysis:
##     For U.S.-based data scientists:
##     average salary: $148297.09 median salary: $138475.00 standard deviation $66655.66 For Offshore-ba
##     average salary: $69529.92 median salary: $63760.50 standard deviation: $43083.56
```

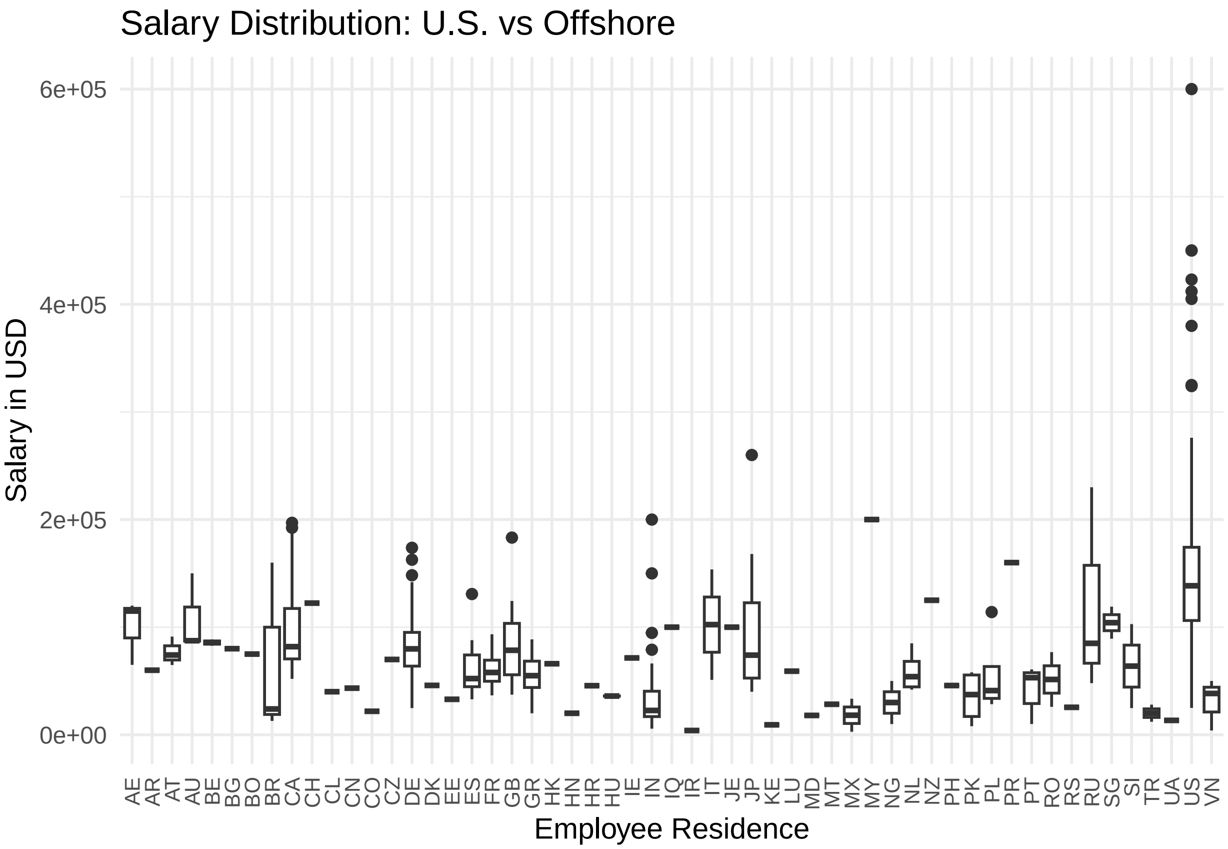
Boxplot for Salary Distribution

There is a wide variation in salary distributions across different residences. Some residences have higher median salaries compared to others. The U.S. shows a relatively high median salary and a wide range of salaries, including several outliers. Other countries also show significant variations, with some having more outliers indicating higher salaries. Disparities in salary distributions could be influenced by factors such as cost of living, demand for skills, and economic conditions in different regions. Nonetheless, as tool for analyzing and understanding global salary trends, this visualization provides a clear comparison of salary distributions across various employee residences.

X-axis: Employee Residence (various country codes) Y-axis: Salary in USD (ranging from 0 to 600,000 USD) Boxes: Represent the interquartile range (IQR), which contains the middle 50% of the data. Horizontal Line Inside the Box: Represents the median salary for that residence. Whiskers: Extend from the boxes to the smallest and largest values within 1.5 times the IQR from the quartiles. Dots: Represent outliers, which are data points outside the whiskers.

```r
# Boxplot for Salary Distribution with adjusted margins
ggplot(full_time_data, aes(x = employee_residence, y = salary_in_usd)) +
  geom_boxplot() +
  theme_minimal() +
  labs(title = "Salary Distribution: U.S. vs Offshore",
       x = "Employee Residence",
       y = "Salary in USD") +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1, size = 8)) +
  theme(plot.margin = unit(c(1,1,1,2), "cm")) %>%
  print()
```

```
## List of 1
##  $ plot.margin: 'simpleUnit' num [1:4] 1cm 1cm 1cm 2cm
##   ..- attr(*, "unit")= int 1
##  - attr(*, "class")= chr [1:2] "theme" "gg"
##  - attr(*, "complete")= logi FALSE
##  - attr(*, "validate")= logi TRUE
```



Bar graph for Average Salary by Experience Level (U.S. versus Offshore)

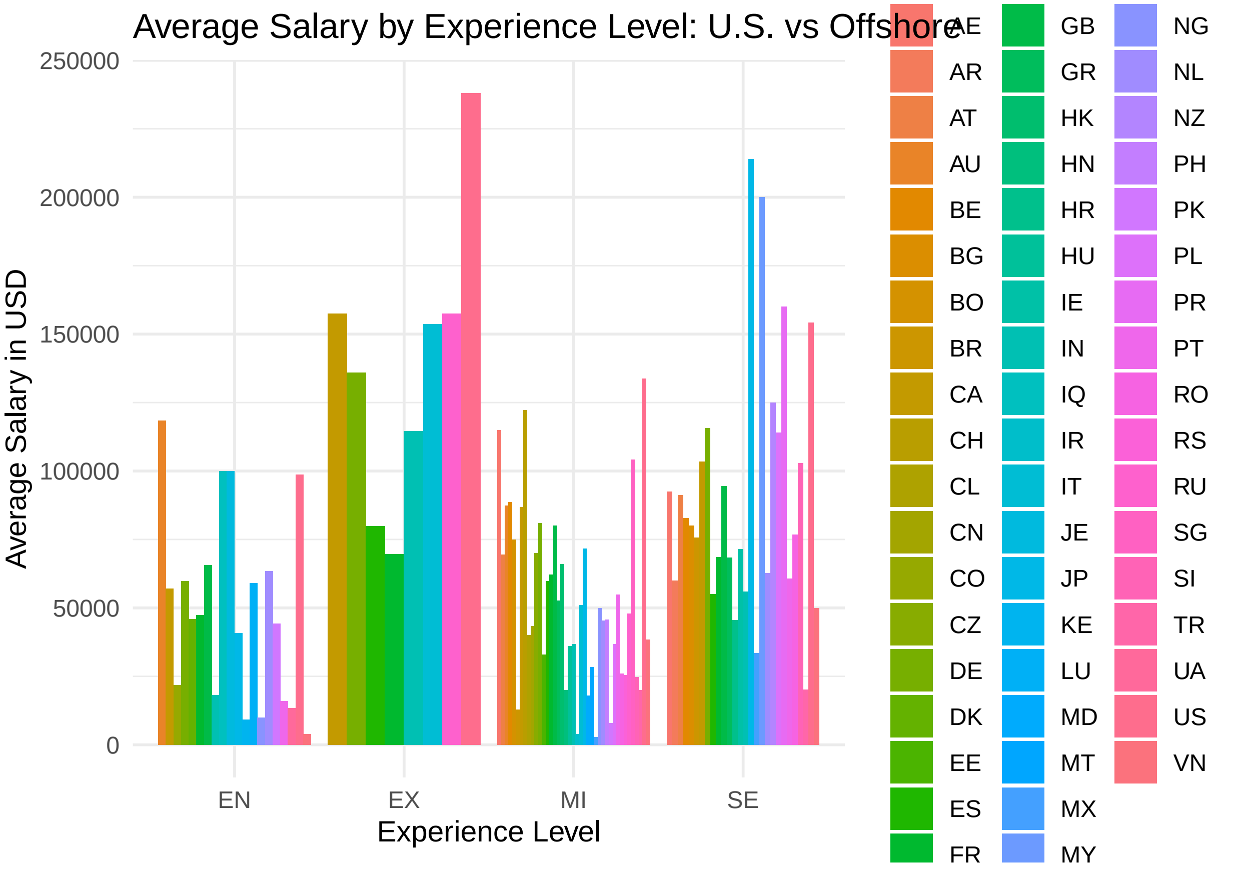
The bar graph provides a clear visual comparison of average salaries for different experience levels across U.S. and Offshore locations. U.S.-based employees have higher average salaries compared to their offshore counterparts across all experience levels. This highlights potential salary gaps and can inform strategic decisions regarding competitive compensation offers by emphasizing the importance of considering geographic location and experience level in salary planning. X-Axis: Experience Level (with categories: Entry Level, Mid-Level, Senior Level, Executive Level) Y-Axis: Average Salary in USD

```r
# Bar plot for Average Salary by Experience Level in U.S. and Offshore
avg_salary_experience <- full_time_data %>%
  group_by(employee_residence, experience_level) %>%
  summarize(avg_salary = mean(salary_in_usd, na.rm = TRUE)) %>%
  ungroup()
```

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

```r
ggplot(avg_salary_experience, aes(x = experience_level, y = avg_salary, fill = employee_residence)) +
  geom_bar(stat = "identity", position = "dodge") +
  theme_minimal() +
  labs(title = "Average Salary by Experience Level: U.S. vs Offshore",
       x = "Experience Level",
       y = "Average Salary in USD",
       fill = "Employee Residence") +
  theme(plot.title = element_text(size = 10, hjust = 0.5)) %>%
  print()
```

```
## List of 1
##  $ plot.title:List of 11
##   ..$ family       : NULL
##   ..$ face         : NULL
##   ..$ colour       : NULL
##   ..$ size         : num 10
##   ..$ hjust        : num 0.5
##   ..$ vjust        : NULL
##   ..$ angle        : NULL
##   ..$ lineheight   : NULL
##   ..$ margin       : NULL
##   ..$ debug        : NULL
##   ..$ inherit.blank: logi FALSE
##   ..- attr(*, "class")= chr [1:2] "element_text" "element"
##  - attr(*, "class")= chr [1:2] "theme" "gg"
##  - attr(*, "complete")= logi FALSE
##  - attr(*, "validate")= logi TRUE
```



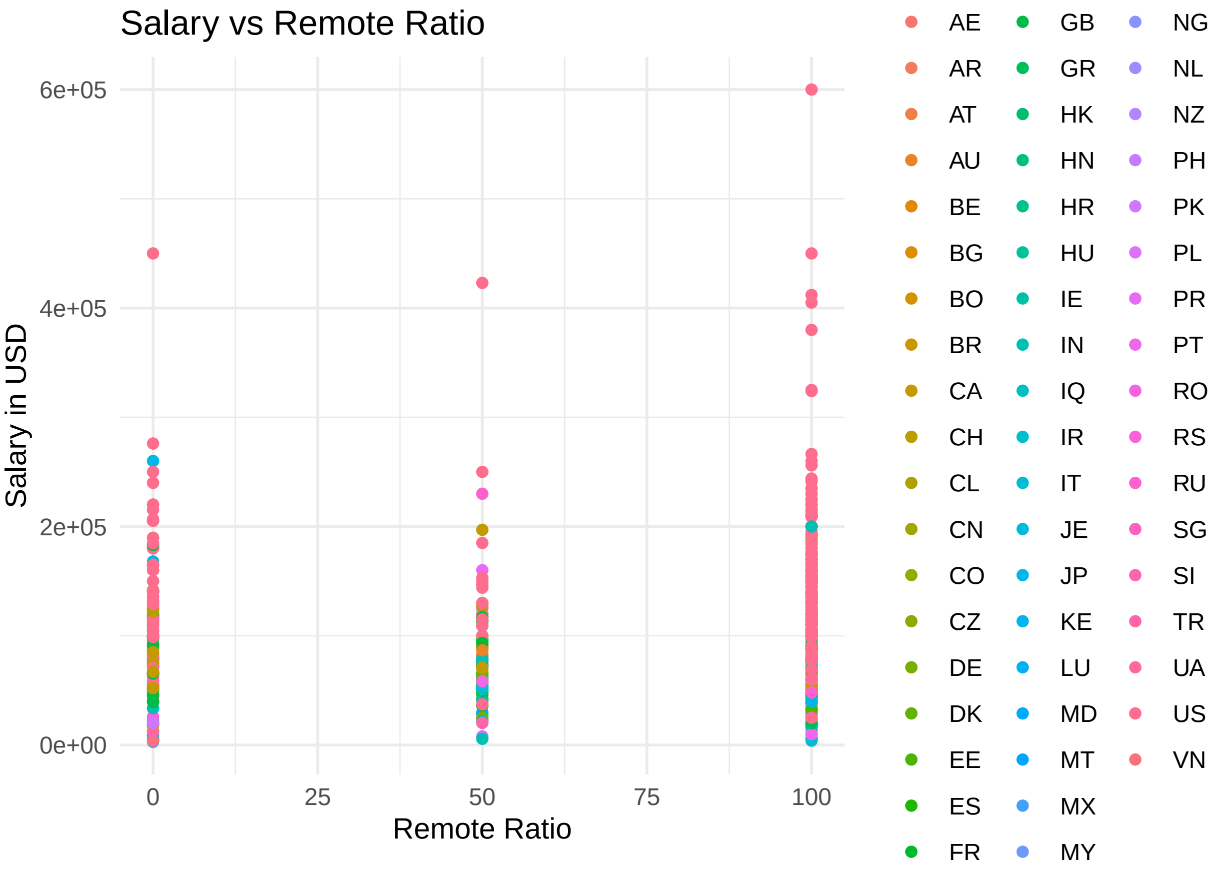
Scatter plot for Salary versus Remote Work Ratio

The scatter plot represents the relationship between salary in USD and remote work ratio for employees from different countries, emphasizing the salary distribution. Salaries vary widely across all remote ratios. Some employees in the fully remote category (100% remote ratio) are earning higher salaries, indicating that remote work may be associated with more competitive compensation for certain roles in in certain regions.

X-Axis (Remote Ratio): This axis represents the percentage of work that employees perform remotely, ranging from 0% (fully on-site) to 100% (fully remote). Y-Axis (Salary in USD): This axis shows the salaries of the employees, measured in US dollars. Color (Employee Residence): Each color represents a different country where the employees reside.

```r
# Scatter plot for Salary versus Remote Work Ratio
ggplot(full_time_data, aes(x = remote_ratio, y = salary_in_usd, color = employee_residence)) +
  geom_point() +
  theme_minimal() +
  labs(title = "Salary vs Remote Ratio",
       x = "Remote Ratio",
       y = "Salary in USD",
       color = "Employee Residence") %>%
  print()
```

```
## $x
## [1] "Remote Ratio"
##
## $y
## [1] "Salary in USD"
##
## $colour
## [1] "Employee Residence"
##
## $title
## [1] "Salary vs Remote Ratio"
##
## attr(,"class")
## [1] "labels"
```



Strategic Salary Recommendations

Balancing U.S. and Offshore Data Science Compensation in a Competitive Market

```
# Conclusion based on the analysis
cat("Based on the analysis, I recommend the following:
    For U.S.-based data scientists,
    the average salary is around", us_summary$avg_salary,
    "USD,
    while offshore data scientists have an average
    salary of approximately", offshore_summary$avg_salary,"USD.
    To attract top talent, especially in a competitive market,
    consider offering salaries at or above these averages
    depending on the candidate's experience and expertise.
    Considering a notable relationship between the remote
    work ratio and salary distribution for employees across
    different countries, as well as, current trends in the
    influence of remote work on hiring salary, more
    competitive compensation may be necessary for roles
    allowing full remote work.

    These considerations may help set competitive salaries
    based on average salaries, remote work ratios,
    and geographical locations. It can also inform decisions
    about remote work policies and assist in developing
    guidelines and policies that support fair compensation
```

```r
    practices.")
```

```
## Based on the analysis, I recommend the following:
##     For U.S.-based data scientists,
##     the average salary is around $148297.09 USD,
##     while offshore data scientists have an average
##     salary of approximately $69529.92 USD.
##     To attract top talent, especially in a competitive market,
##     consider offering salaries at or above these averages
##     depending on the candidate's experience and expertise.
##     Considering a notable relationship between the remote
##     work ratio and salary distribution for employees across
##     different countries, as well as, current trends in the
##     influence of remote work on hiring salary, more
##     competitive compensation may be necessary for roles
##     allowing full remote work.
##
##     These considerations may help set competitive salaries
##     based on average salaries, remote work ratios,
##     and geographical locations. It can also inform decisions
##     about remote work policies and assist in developing
##     guidelines and policies that support fair compensation
##     practices.
```

```r
company_data <- data %>% filter(employment_type == "FT")
# Summarize the full_time_data dataframe
summary(full_time_data)
```

```
##       ...1         work_year    experience_level   employment_type
##   Min.   :  0.0   Min.   :2020   Length:588         Length:588
##   1st Qu.:155.8   1st Qu.:2021   Class :character   Class :character
##   Median :308.5   Median :2022   Mode  :character   Mode  :character
##   Mean   :306.0   Mean   :2021
##   3rd Qu.:455.2   3rd Qu.:2022
##   Max.   :606.0   Max.   :2022
##    job_title              salary          salary_currency    salary_in_usd
##   Length:588          Min.   :    4000   Length:588         Min.   :  2859
##   Class :character   1st Qu.:   70000   Class :character   1st Qu.: 64962
##   Mode  :character   Median :  115250   Mode  :character   Median :104196
##                      Mean   :  331125                      Mean   :113468
##                      3rd Qu.:  165000                      3rd Qu.:150000
##                      Max.   :30400000                      Max.   :600000
##   employee_residence  remote_ratio     company_location   company_size
##   Length:588          Min.   :  0.00   Length:588         Length:588
##   Class :character   1st Qu.: 50.00   Class :character   Class :character
##   Mode  :character   Median :100.00   Mode  :character   Mode  :character
##                      Mean   : 70.75
##                      3rd Qu.:100.00
##                      Max.   :100.00
```

```r
# Display the first few rows of the full_time_data dataframe
head(full_time_data)
```

```
## # A tibble: 6 x 12
##    ...1 work_year experience_level employment_type job_title              salary
##   <dbl>     <dbl> <chr>            <chr>           <chr>                    <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 <chr>, salary_in_usd <dbl>,
## #   employee_residence <chr>, remote_ratio <dbl>, company_location <chr>,
## #   company_size <chr>
```

```
# List the column names of the full_time_data dataframe
colnames(full_time_data)
```

```
##  [1] "...1"               "work_year"          "experience_level"
##  [4] "employment_type"    "job_title"          "salary"
##  [7] "salary_currency"    "salary_in_usd"      "employee_residence"
## [10] "remote_ratio"       "company_location"   "company_size"
```

```
# Filter data for full-time employees and calculate the averages
average_remote_work <- company_data %>%
  filter(employment_type == "FT") %>%
  group_by(company_location, job_title, remote_ratio) %>%
  summarise(average_percentage = mean(as.numeric(remote_ratio), na.rm = TRUE) * 100) %>%
  ungroup()
```

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

```
# Filter data for full-time employees and calculate the averages
average_remote_work <- company_data %>%
  filter(employment_type == "FT") %>%
  mutate(group = ifelse(company_location == "US", "US", "Offshore")) %>%
  group_by(group, company_size, remote_ratio) %>%
  summarise(average_percentage = n() / nrow(company_data) * 100) %>%
  ungroup()
```

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

```
# Reshape the data for better readability
average_remote_work_table <- average_remote_work %>%
  pivot_wider(names_from = remote_ratio, values_from = average_percentage, names_prefix = "Remote_") %>
  rename(Remote_0 = `Remote_0`, Remote_50 = `Remote_50`, Remote_100 = `Remote_100`) %>%
  arrange(group, company_size)
```

Average Percentage of Full-Time Employees Working Remotely by Company Size and Group (US versus Offshore)

The data represented in chart indicates that US companies, especially medium-sized ones, have a higher percentage of full-time employees working remotely compared to Offshore companies.

Remote work trends across different company sizes and geographical locations highlight the variations in the adoption of remote work between large, medium, and small companies in the US and Offshore.

```r
# Display the data as a table
kable(average_remote_work_table, caption = "Average Percentage of Full-Time Employees Working
      Remotely by Company Size and Group (US/Offshore)",
      col.names = c("Group", "Company Size", "0% Remote", "50% Remote", "100% Remote"))
```

Table 1: Average Percentage of Full-Time Employees Working Remotely by Company Size and Group (US/Offshore)

Group	Company Size	0% Remote	50% Remote	100% Remote
Offshore	L	2.0408163	7.4829932	5.782313
Offshore	M	5.7823129	2.7210884	9.013605
Offshore	S	2.0408163	2.0408163	4.251701
US	L	3.2312925	2.5510204	11.734694
US	M	7.6530612	0.1700680	28.741497
US	S	0.6802721	0.6802721	3.401361

```r
#{r echo=FALSE, message=FALSE, warning=FALSE, results='hide'}
```