Project1

Mike Steyer

2024-11-23

```
paletteName <- "Dark2"
```

```
## read in the data file from the current dir's "r project data.csv" file
salary_df <- read_csv("./r project data.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`
```

```
## factorize where appropriate
salary_df$experience_level <- factor(salary_df$experience_level, levels=c("EN","MI","SE","EX"))
salary_df$employment_type <- factor(salary_df$employment_type)
salary_df$salary_currency <- factor(salary_df$salary_currency)
salary_df$employee_residence <- factor(salary_df$employee_residence)
salary_df$company_location <- factor(salary_df$company_location)
salary_df$company_size <- factor(salary_df$company_size)
salary_df$work_year <- factor(salary_df$work_year)

## add a column that cleans up the remote work var
salary_df <- salary_df %>%
  mutate(remote_type = case_when(
    remote_ratio == 0 ~ "None",
    remote_ratio == 50 ~ "Hybrid",
    remote_ratio == 100 ~ "Remote"
    )
  ) %>%
  select(!c(remote_ratio))

salary_df$remote_type <- factor(salary_df$remote_type)

head(salary_df)
```

```
## # A tibble: 6 × 12
##    ...1 work_year experience_level employment_type job_title              salary
##   <dbl> <fct>     <fct>            <fct>           <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 <fct>, salary_in_usd <dbl>,
## #   employee_residence <fct>, company_location <fct>, company_size <fct>,
## #   remote_type <fct>
```

```
str(salary_df)
```

```
## tibble [607 × 12] (S3: tbl_df/tbl/data.frame)
##  $ ...1              : num [1:607] 0 1 2 3 4 5 6 7 8 9 ...
##  $ work_year         : Factor w/ 3 levels "2020","2021",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ experience_level  : Factor w/ 4 levels "EN","MI","SE",..: 2 3 3 2 3 1 3 2 2 3 ...
##  $ employment_type   : Factor w/ 4 levels "CT","FL","FT",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ job_title         : chr [1:607] "Data Scientist" "Machine Learning Scientist" "Big Data En
gineer" "Product Data Analyst" ...
##  $ salary            : num [1:607] 70000 260000 85000 20000 150000 72000 190000 11000000 1350
00 125000 ...
##  $ salary_currency   : Factor w/ 17 levels "AUD","BRL","CAD",..: 8 17 9 17 17 17 17 10 17 17
...
##  $ salary_in_usd     : num [1:607] 79833 260000 109024 20000 150000 ...
##  $ employee_residence: Factor w/ 57 levels "AE","AR","AT",..: 15 33 21 24 56 56 56 26 56 42
...
##  $ company_location  : Factor w/ 50 levels "AE","AS","AT",..: 13 30 19 21 49 49 49 23 49 39
...
##  $ company_size      : Factor w/ 3 levels "L","M","S": 1 3 2 3 1 1 3 1 1 3 ...
##  $ remote_type       : Factor w/ 3 levels "Hybrid","None",..: 2 2 1 2 1 3 3 1 3 1 ...
```

```
## Let's analyze the titles and try to get them as close to comparable as possible
unique(salary_df$job_title)
```

```
##  [1] "Data Scientist"
##  [2] "Machine Learning Scientist"
##  [3] "Big Data Engineer"
##  [4] "Product Data Analyst"
##  [5] "Machine Learning Engineer"
##  [6] "Data Analyst"
##  [7] "Lead Data Scientist"
##  [8] "Business Data Analyst"
##  [9] "Lead Data Engineer"
## [10] "Lead Data Analyst"
## [11] "Data Engineer"
## [12] "Data Science Consultant"
## [13] "BI Data Analyst"
## [14] "Director of Data Science"
## [15] "Research Scientist"
## [16] "Machine Learning Manager"
## [17] "Data Engineering Manager"
## [18] "Machine Learning Infrastructure Engineer"
## [19] "ML Engineer"
## [20] "AI Scientist"
## [21] "Computer Vision Engineer"
## [22] "Principal Data Scientist"
## [23] "Data Science Manager"
## [24] "Head of Data"
## [25] "3D Computer Vision Researcher"
## [26] "Data Analytics Engineer"
## [27] "Applied Data Scientist"
## [28] "Marketing Data Analyst"
## [29] "Cloud Data Engineer"
## [30] "Financial Data Analyst"
## [31] "Computer Vision Software Engineer"
## [32] "Director of Data Engineering"
## [33] "Data Science Engineer"
## [34] "Principal Data Engineer"
## [35] "Machine Learning Developer"
## [36] "Applied Machine Learning Scientist"
## [37] "Data Analytics Manager"
## [38] "Head of Data Science"
## [39] "Data Specialist"
## [40] "Data Architect"
## [41] "Finance Data Analyst"
## [42] "Principal Data Analyst"
## [43] "Big Data Architect"
## [44] "Staff Data Scientist"
## [45] "Analytics Engineer"
## [46] "ETL Developer"
## [47] "Head of Machine Learning"
## [48] "NLP Engineer"
## [49] "Lead Machine Learning Engineer"
## [50] "Data Analytics Lead"
```

```
## Not too much to lean on for job_title, let's just use the experience level attached to the re
cord
##    (job titles are notoriously unhelpful when it comes to judging competence anyway)
## Create a flag indicating if they're "leadership potential" - senior or expert/director
leadership_potential_experience_levels = c("SE","EX")

salary_df <- salary_df %>%
  mutate(is_leadership_potential = experience_level %in% leadership_potential_experience_levels)

## create flags indicating domestic (US) vs offshore (not US)
salary_df <- salary_df %>%
  mutate(employee_country_type = ifelse(employee_residence == "US", "Domestic", "Offshore")) %>%
  mutate(company_country_type = ifelse(company_location == "US", "Domestic", "Offshore")) %>%
  mutate(employment_status = paste("Company ", company_country_type, ", Employee ", employee_cou
ntry_type, sep="")) %>%
  # only include full-time
  filter(employment_type == "FT")

salary_df$employee_country_type <- factor(salary_df$employee_country_type)
salary_df$company_country_type <- factor(salary_df$company_country_type)
salary_df$employment_status <- factor(salary_df$employment_status)

head(salary_df)
```

```
## # A tibble: 6 × 16
##    ...1 work_year experience_level employment_type job_title              salary
##   <dbl> <fct>     <fct>            <fct>           <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 10 more variables: salary_currency <fct>, salary_in_usd <dbl>,
## #   employee_residence <fct>, company_location <fct>, company_size <fct>,
## #   remote_type <fct>, is_leadership_potential <lgl>,
## #   employee_country_type <fct>, company_country_type <fct>,
## #   employment_status <fct>
```

```
## let's shoot for the following:
## - give an introduction of the data set and the different facets
##    - total number of US based company positions
##    - pie charts showing the year vs total
## - show the trend of median "data job" salaries over time, stats of latest year (median, mean,
min, max) - US companies only
##    - US employee vs non-US employee
## - show the trend of median leadership salaries over time, stats of latest year (median, mean,
min, max) - US companies only
##    - US employee vs non-US employee
## - show median salary by company size by experience level for latest year (US employees)
## - show median salary by company size by experience level for latest year (offshore employees)
## - show salaries by office type (boxplot)
```

```
summary(salary_df)
```

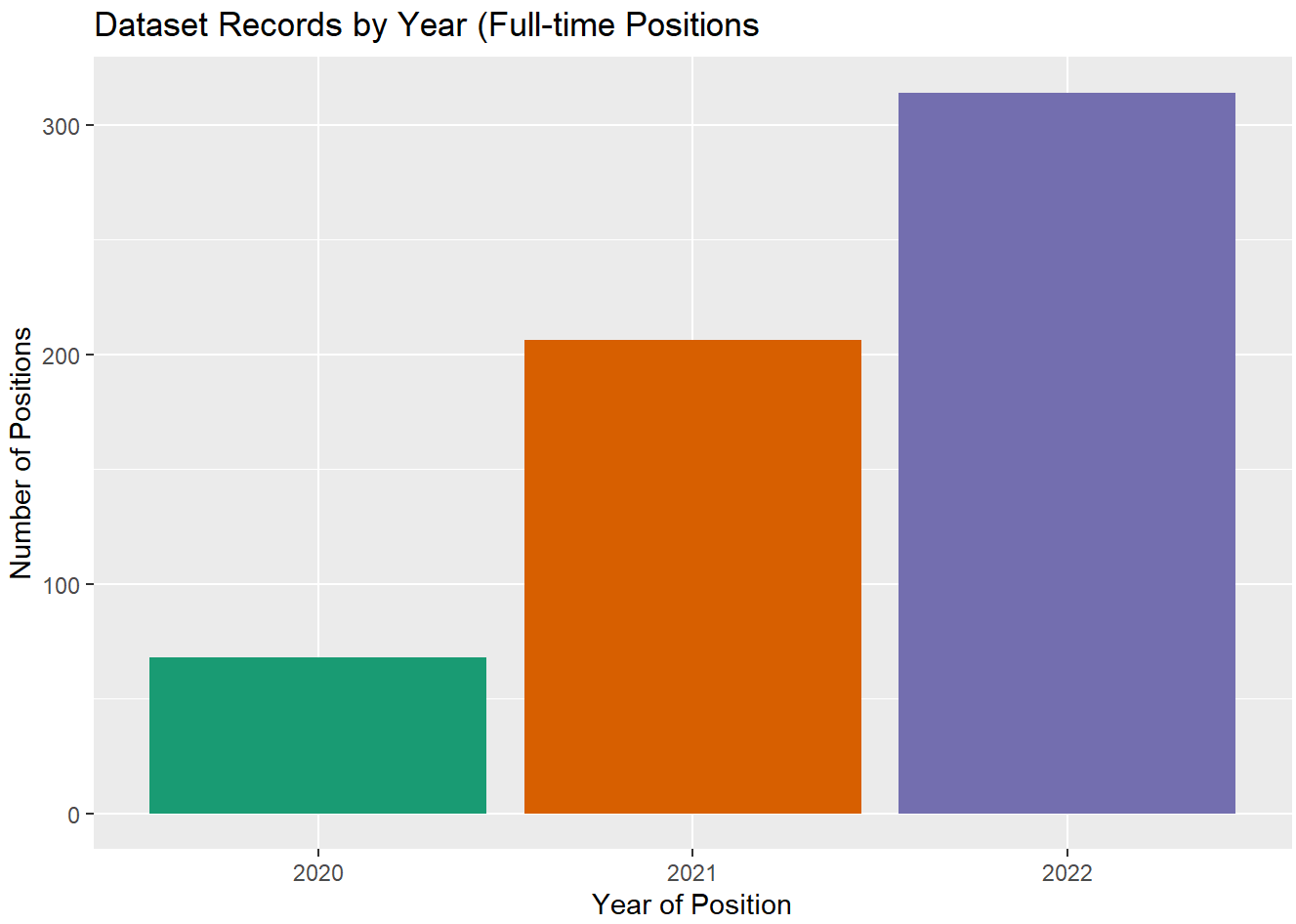
```
##       ...1       work_year  experience_level employment_type  job_title
##  Min.   :  0.0   2020: 68   EN: 79           CT:  0          Length:588
##  1st Qu.:155.8   2021:206   MI:206           FL:  0          Class :character
##  Median :308.5   2022:314   SE:278           FT:588          Mode  :character
##  Mean   :306.0              EX: 25           PT:  0
##  3rd Qu.:455.2
##  Max.   :606.0
##
##      salary         salary_currency salary_in_usd    employee_residence
##  Min.   :    4000   USD    :387     Min.   :  2859   US     :328
##  1st Qu.:   70000   EUR    : 89     1st Qu.: 64962   GB     : 44
##  Median :  115250   GBP    : 44     Median :104197   IN     : 29
##  Mean   :  331125   INR    : 26     Mean   :113468   CA     : 28
##  3rd Qu.:  165000   CAD    : 18     3rd Qu.:150000   DE     : 23
##  Max.   :30400000   JPY    :  3     Max.   :600000   FR     : 18
##                     (Other): 21                      (Other):118
##  company_location company_size remote_type  is_leadership_potential
##  US     :346      L:193        Hybrid: 92   Mode :logical
##  GB     : 47      M:318        None  :126   FALSE:285
##  CA     : 30      S: 77        Remote:370   TRUE :303
##  DE     : 26
##  IN     : 23
##  FR     : 15
##  (Other):101
##  employee_country_type company_country_type
##  Domestic:328          Domestic:346
##  Offshore:260          Offshore:242
##
##
##
##
##
##                            employment_status
##  Company Domestic, Employee Domestic:326
##  Company Domestic, Employee Offshore: 20
##  Company Offshore, Employee Domestic:  2
##  Company Offshore, Employee Offshore:240
##
##
##
```

```
str(salary_df)
```

```
## tibble [588 × 16] (S3: tbl_df/tbl/data.frame)
##  $ ...1                   : num [1:588] 0 1 2 3 4 5 6 7 8 9 ...
##  $ work_year              : Factor w/ 3 levels "2020","2021",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ experience_level       : Factor w/ 4 levels "EN","MI","SE",..: 2 3 3 2 3 1 3 2 2 3 ...
##  $ employment_type        : Factor w/ 4 levels "CT","FL","FT",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ job_title              : chr [1:588] "Data Scientist" "Machine Learning Scientist" "Big Da
ta Engineer" "Product Data Analyst" ...
##  $ salary                 : num [1:588] 70000 260000 85000 20000 150000 72000 190000 11000000
135000 125000 ...
##  $ salary_currency        : Factor w/ 17 levels "AUD","BRL","CAD",..: 8 17 9 17 17 17 17 10 1
7 17 ...
##  $ salary_in_usd          : num [1:588] 79833 260000 109024 20000 150000 ...
##  $ employee_residence     : Factor w/ 57 levels "AE","AR","AT",..: 15 33 21 24 56 56 56 26 56
42 ...
##  $ company_location       : Factor w/ 50 levels "AE","AS","AT",..: 13 30 19 21 49 49 49 23 49
39 ...
##  $ company_size           : Factor w/ 3 levels "L","M","S": 1 3 2 3 1 1 3 1 1 3 ...
##  $ remote_type            : Factor w/ 3 levels "Hybrid","None",..: 2 2 1 2 1 3 3 1 3 1 ...
##  $ is_leadership_potential: logi [1:588] FALSE TRUE TRUE FALSE TRUE FALSE ...
##  $ employee_country_type  : Factor w/ 2 levels "Domestic","Offshore": 2 2 2 2 1 1 1 2 1 2 ...
##  $ company_country_type   : Factor w/ 2 levels "Domestic","Offshore": 2 2 2 2 1 1 1 2 1 2 ...
##  $ employment_status      : Factor w/ 4 levels "Company Domestic, Employee Domestic",..: 4 4
4 4 1 1 1 4 1 4 ...
```

```
## Data Set Overview:
## Total records: 607
## Total US-based company positions: 355
## Salary amounts in USD
```

```
## US Company Positions by Year
## pie chart of the amount of US-based company positions found by year
## src: https://r-graph-gallery.com/piechart-ggplot2.html, https://r-charts.com/part-whole/pie-c
hart-ggplot2/
## output: ggsave()

salary_df %>%
  group_by(work_year) %>%
  summarize(count = n()) %>%
  ggplot(aes(x=work_year, y=count, fill=work_year)) +
    geom_bar(stat="identity") +
    scale_fill_brewer(palette = paletteName) +
    theme(legend.position="none") +
    ggtitle("Dataset Records by Year (Full-time Positions") +
    xlab("Year of Position") +
    ylab("Number of Positions")
```



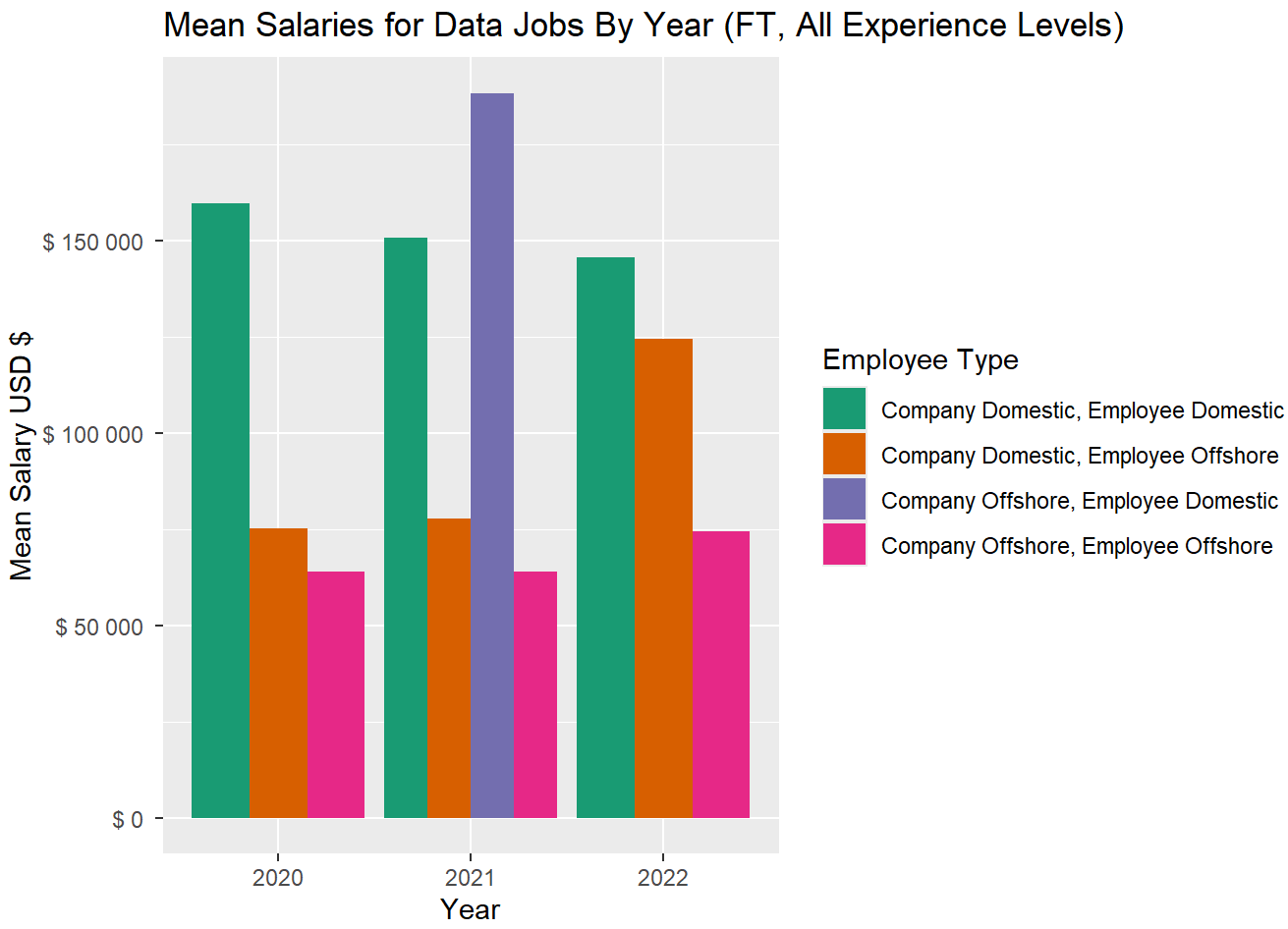

```
ggsave("records-by-year.png")
```

```
## Saving 7 x 5 in image
```

```
## Trend "data job" position median salaries by year for US-based companies (on & offshore)

all_data_jobs_salary_by_year <- salary_df %>%
  group_by(work_year, employment_status) %>%
  summarize(
    median_salary_in_usd = median(salary_in_usd),
    mean_salary_in_usd = mean(salary_in_usd),
    maximum_salary_in_usd = max(salary_in_usd),
    minimum_salary_in_usd = min(salary_in_usd),
    count = n(),
    .groups = "drop"
  )


all_data_jobs_salary_by_year
```

```
## # A tibble: 10 × 7
##    work_year employment_status           median_salary_in_usd mean_salary_in_usd
##    <fct>     <fct>                                    <dbl>              <dbl>
##  1 2020      Company Domestic, Employee…              119000            159856.
##  2 2020      Company Domestic, Employee…               62214             75247
##  3 2020      Company Offshore, Employee…               49724             64084.
##  4 2021      Company Domestic, Employee…              137500            150804.
##  5 2021      Company Domestic, Employee…               54094             77835.
##  6 2021      Company Offshore, Employee…              188500            188500
##  7 2021      Company Offshore, Employee…               61467             64037.
##  8 2022      Company Domestic, Employee…              140000            145736.
##  9 2022      Company Domestic, Employee…              100000            124600
## 10 2022      Company Offshore, Employee…               68147             74452.
## # ℹ 3 more variables: maximum_salary_in_usd <dbl>, minimum_salary_in_usd <dbl>,
## #   count <int>
```

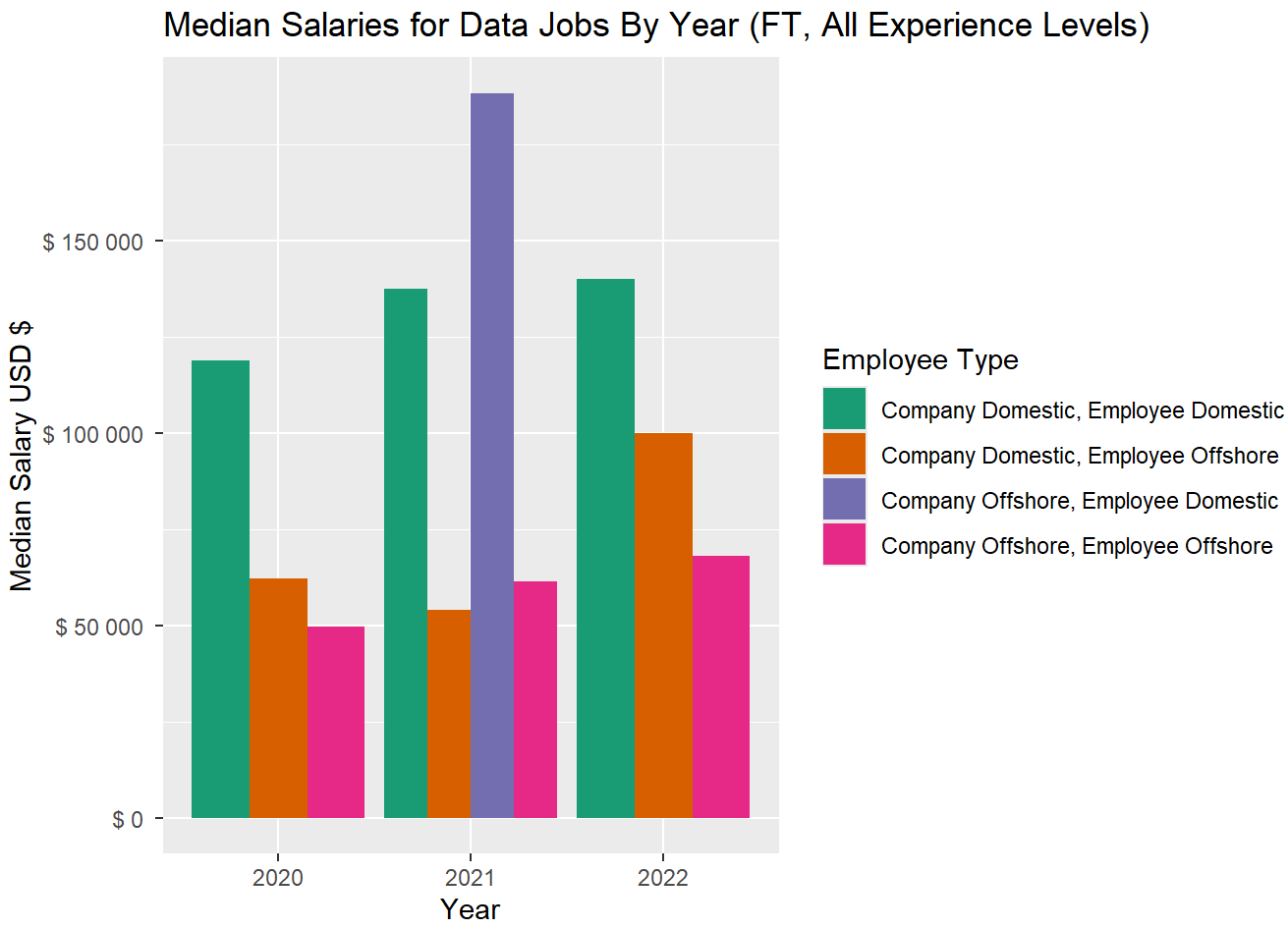
```
all_data_jobs_salary_by_year %>%
    ggplot(aes(x=work_year, y=mean_salary_in_usd, fill=employment_status)) +
    geom_bar(stat="identity", position = "dodge") +
    scale_fill_brewer(palette = paletteName) +
    xlab("Year") +
    ylab("Mean Salary USD $") +
    guides(fill=guide_legend(title="Employee Type")) +
    ggtitle("Mean Salaries for Data Jobs By Year (FT, All Experience Levels)") +
    scale_y_continuous(labels = scales::unit_format(prefix ="$ ", unit=""))
```




```
ggsave("all-data-jobs-salary-by-year.png")
```

```
## Saving 7 x 5 in image
```

```
all_data_jobs_salary_by_year %>%
    ggplot(aes(x=work_year, y=median_salary_in_usd, fill=employment_status)) +
    geom_bar(stat="identity", position = "dodge") +
    scale_fill_brewer(palette = paletteName) +
    xlab("Year") +
    ylab("Median Salary USD $") +
    guides(fill=guide_legend(title="Employee Type")) +
    ggtitle("Median Salaries for Data Jobs By Year (FT, All Experience Levels)") +
    scale_y_continuous(labels = scales::unit_format(prefix ="$ ", unit=""))
```



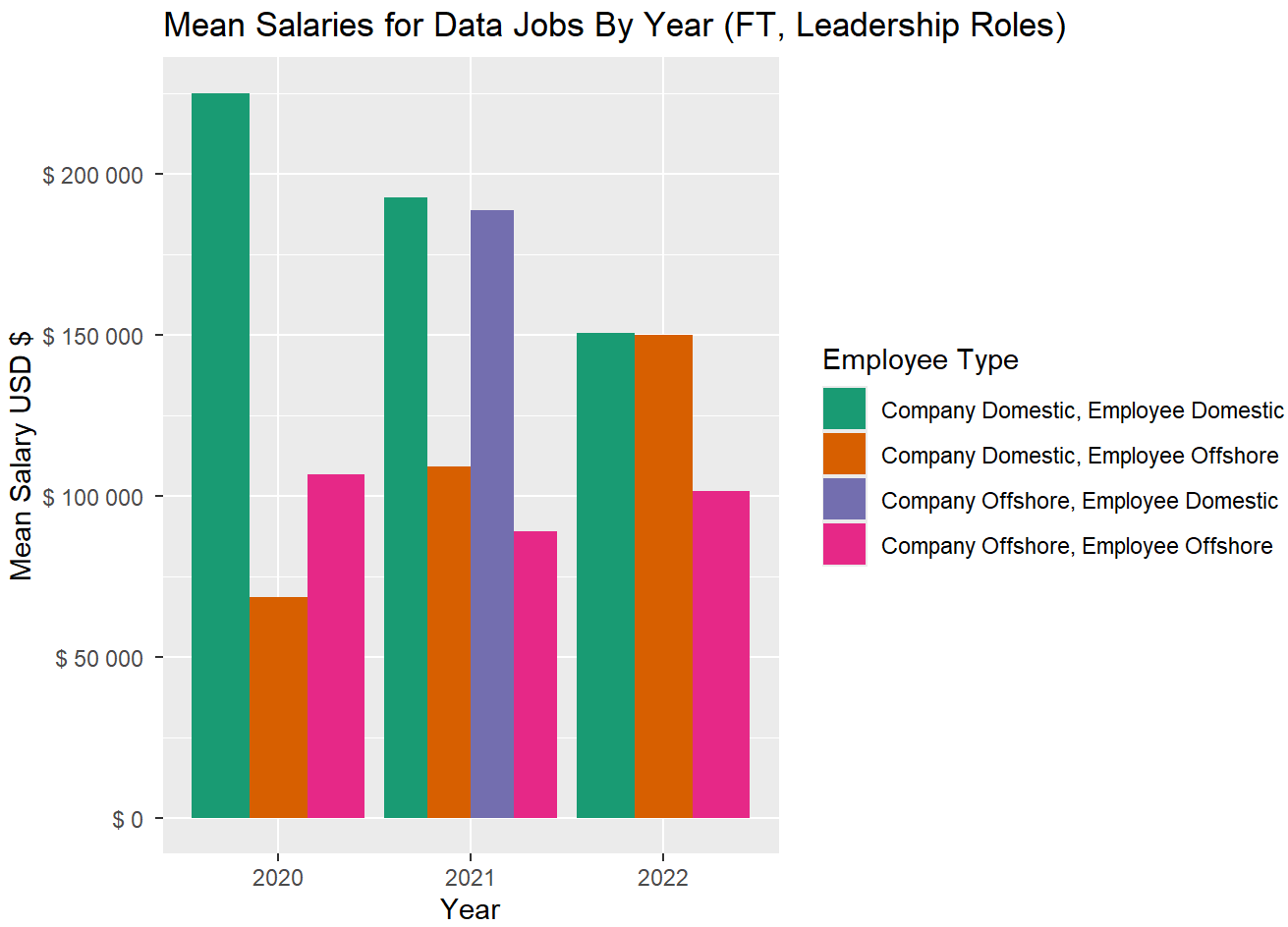

```
ggsave("all-data-jobs-salary-by-year-median.png")
```

```
## Saving 7 x 5 in image
```

```
## - show the trend of median leadership salaries over time, stats of latest year (median, mean, min, max) - US companies only
##    - US employee vs non-US employee
leadership_data_jobs_salary_by_year <- salary_df %>%
  filter(is_leadership_potential == TRUE) %>%
  group_by(work_year, employment_status) %>%
  summarize(
    median_salary_in_usd = median(salary_in_usd),
    mean_salary_in_usd = mean(salary_in_usd),
    maximum_salary_in_usd = max(salary_in_usd),
    minimum_salary_in_usd = min(salary_in_usd),
    count = n(),
    .groups = "drop"
  )

leadership_data_jobs_salary_by_year
```

```
## # A tibble: 10 × 7
##    work_year employment_status           median_salary_in_usd mean_salary_in_usd
##    <fct>     <fct>                                    <dbl>              <dbl>
##  1 2020      Company Domestic, Employee…              190000            225029.
##  2 2020      Company Domestic, Employee…               68428             68428 
##  3 2020      Company Offshore, Employee…              109024            106503.
##  4 2021      Company Domestic, Employee…              174000            192636.
##  5 2021      Company Domestic, Employee…              115000            109016.
##  6 2021      Company Offshore, Employee…              188500            188500 
##  7 2021      Company Offshore, Employee…               85000             88847.
##  8 2022      Company Domestic, Employee…              140400            150534.
##  9 2022      Company Domestic, Employee…              150000            150000 
## 10 2022      Company Offshore, Employee…               89316            101285.
## # ℹ 3 more variables: maximum_salary_in_usd <dbl>, minimum_salary_in_usd <dbl>,
## #   count <int>
```

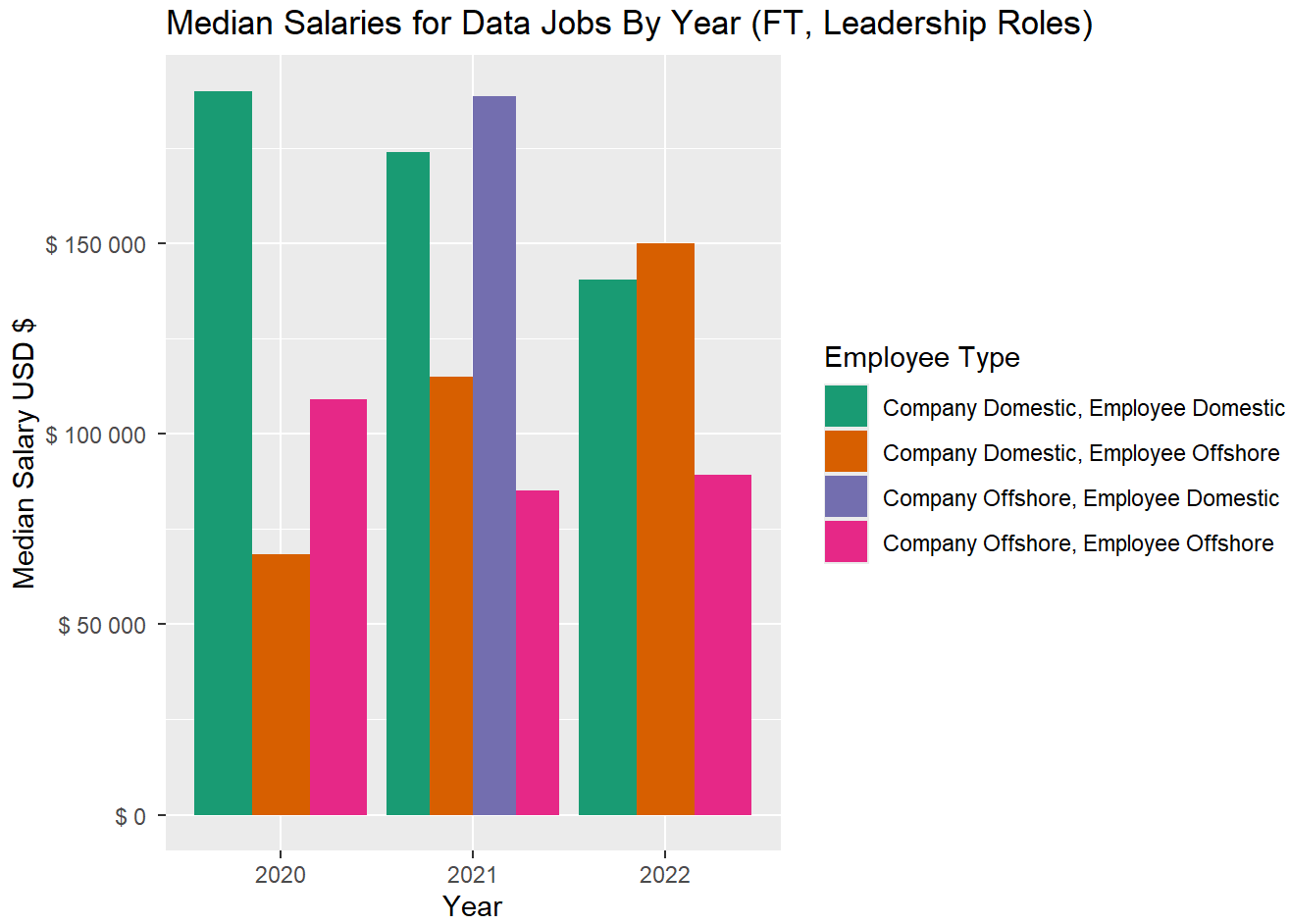
```
leadership_data_jobs_salary_by_year %>%
    ggplot(aes(x=work_year, y=mean_salary_in_usd, fill=employment_status)) +
    geom_bar(stat = "identity", position = "dodge") +
    scale_fill_brewer(palette = paletteName) +
    xlab("Year") +
    ylab("Mean Salary USD $") +
    guides(fill=guide_legend(title="Employee Type")) +
    ggtitle("Mean Salaries for Data Jobs By Year (FT, Leadership Roles)") +
    scale_y_continuous(labels = scales::unit_format(prefix ="$ ", unit=""))
```




```
ggsave("leadership-data-jobs-salary-by-year.png")
```

```
## Saving 7 x 5 in image
```

```
leadership_data_jobs_salary_by_year %>%
    ggplot(aes(x=work_year, y=median_salary_in_usd, fill=employment_status)) +
    geom_bar(stat = "identity", position = "dodge") +
    scale_fill_brewer(palette = paletteName) +
    xlab("Year") +
    ylab("Median Salary USD $") +
    guides(fill=guide_legend(title="Employee Type")) +
    ggtitle("Median Salaries for Data Jobs By Year (FT, Leadership Roles)") +
    scale_y_continuous(labels = scales::unit_format(prefix ="$ ", unit=""))
```



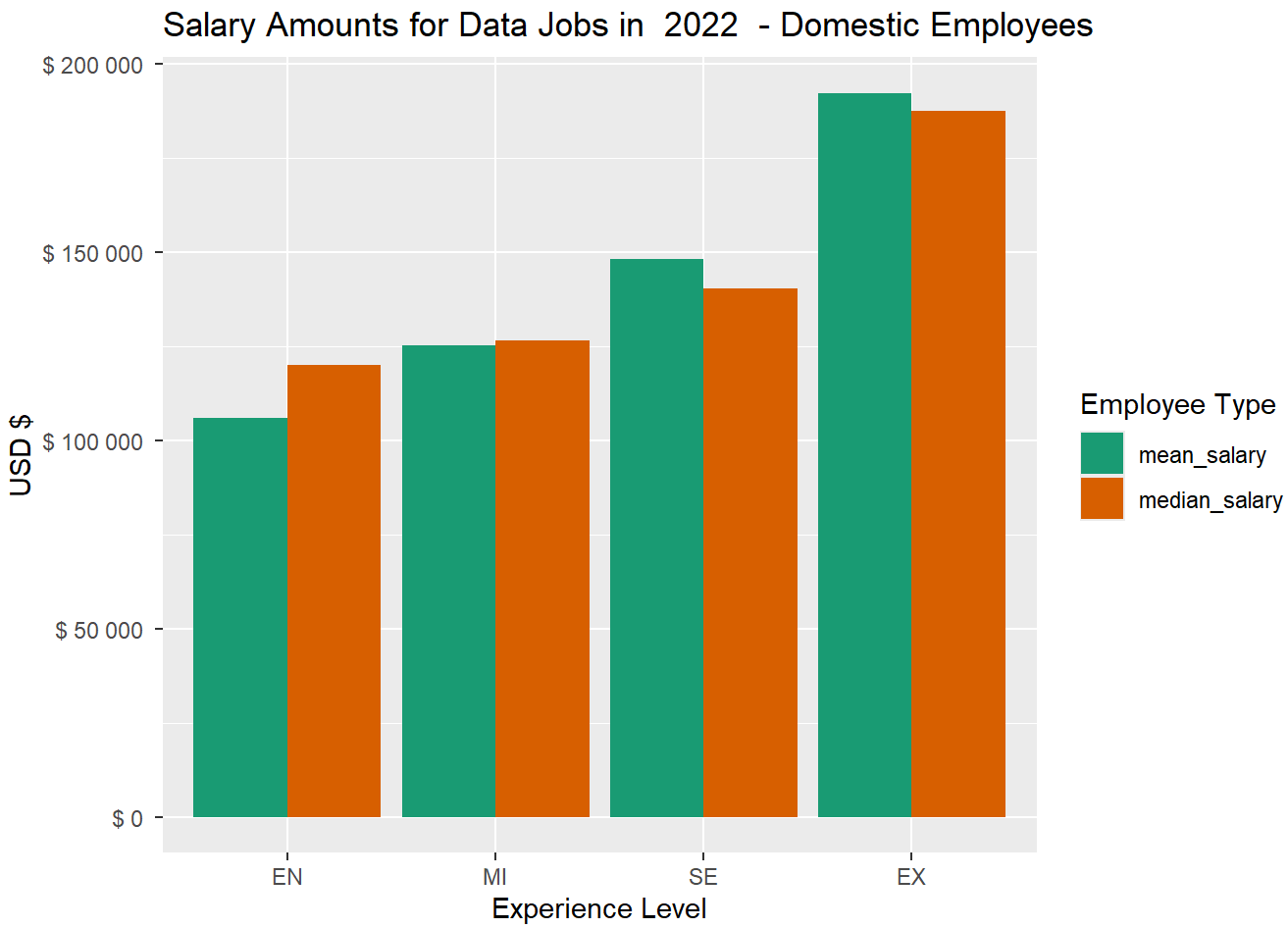

```
ggsave("leadership-data-jobs-salary-by-year-median.png")
```

```
## Saving 7 x 5 in image
```

```
##

latest_year <- 2022

salary_latest_full_time_us_companies_df <- salary_df %>%
  filter(company_country_type == "Domestic") %>%
  filter(work_year == latest_year & employment_type == "FT")

salary_latest_full_time_us_companies_long_df <- pivot_longer(
      salary_latest_full_time_us_companies_df %>%
      group_by(experience_level, employee_country_type) %>%
      summarize(
        median_salary = median(salary_in_usd),
        mean_salary = mean(salary_in_usd),
        .groups = "drop"
      ),
      !c(experience_level, employee_country_type),
      names_to = "metric",
      values_to = "value"
    )

salary_latest_full_time_us_companies_long_df
```

```
## # A tibble: 12 × 4
##    experience_level employee_country_type metric          value
##    <fct>            <fct>                 <chr>           <dbl>
##  1 EN               Domestic              median_salary 120000
##  2 EN               Domestic              mean_salary   106000
##  3 MI               Domestic              median_salary 126500
##  4 MI               Domestic              mean_salary   125297.
##  5 MI               Offshore              median_salary  75000
##  6 MI               Offshore              mean_salary   107667.
##  7 SE               Domestic              median_salary 140325
##  8 SE               Domestic              mean_salary   148101.
##  9 SE               Offshore              median_salary 150000
## 10 SE               Offshore              mean_salary   150000
## 11 EX               Domestic              median_salary 187500
## 12 EX               Domestic              mean_salary   192388.
```

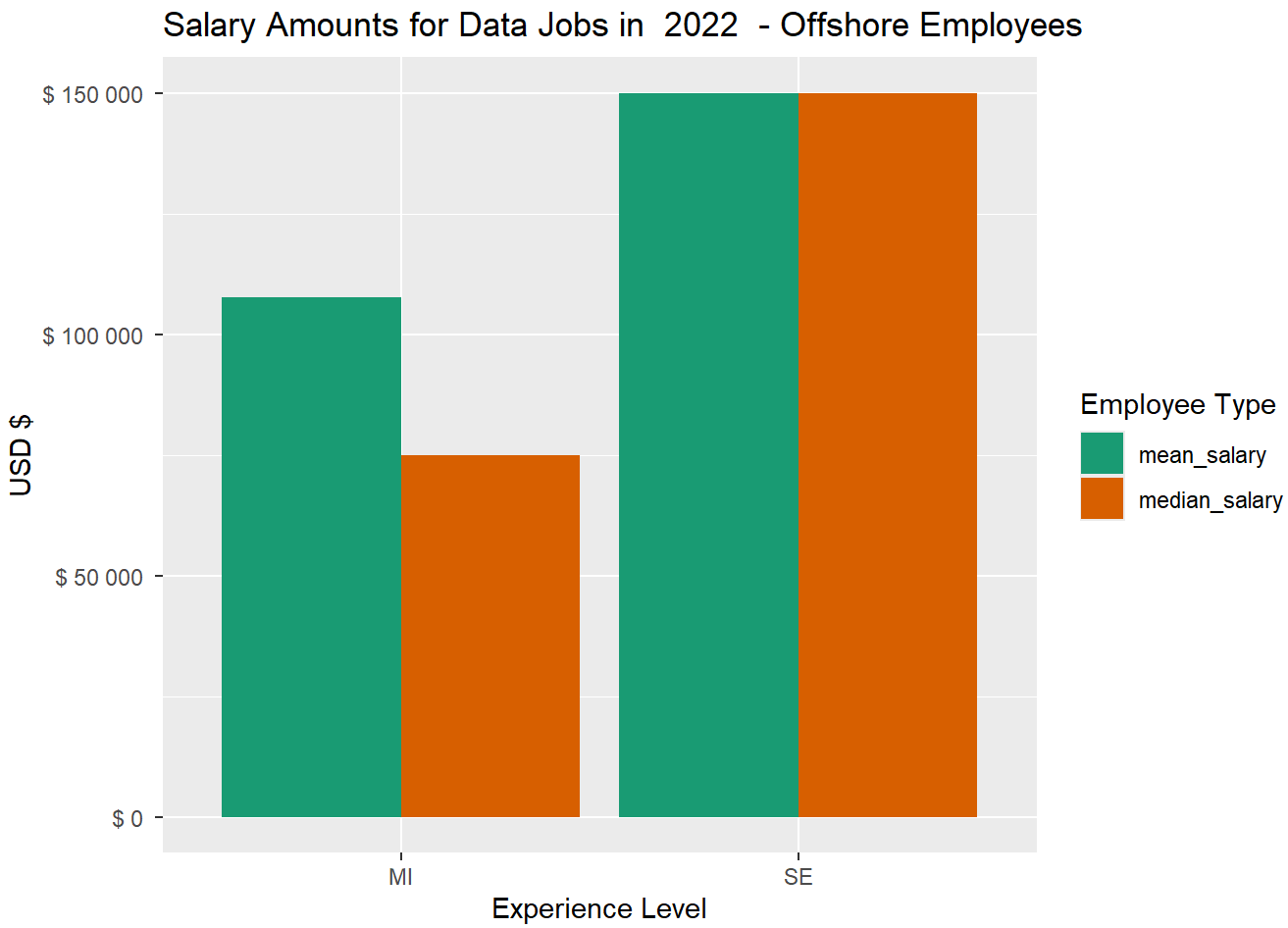
```
salary_latest_full_time_us_companies_long_df %>%
  filter(employee_country_type == "Domestic") %>%
  ggplot(aes(x= experience_level, y=value, fill=metric)) +
    geom_bar(stat="identity", position = "dodge") +
    scale_fill_brewer(palette = paletteName) +
    xlab("Experience Level") +
    ylab("USD $") +
    guides(fill=guide_legend(title="Employee Type")) +
    ggtitle(paste("Salary Amounts for Data Jobs in ", latest_year, " - Domestic Employees")) +
    scale_y_continuous(labels = scales::unit_format(prefix ="$ ", unit=""))
```




```
ggsave("domestic-salaries-2022-by-experience.png")
```

```
## Saving 7 x 5 in image
```

```
salary_latest_full_time_us_companies_long_df %>%
  filter(employee_country_type == "Offshore") %>%
  ggplot(aes(x= experience_level, y=value, fill=metric)) +
    geom_bar(stat="identity", position = "dodge") +
    scale_fill_brewer(palette = paletteName) +
    xlab("Experience Level") +
    ylab("USD $") +
    guides(fill=guide_legend(title="Employee Type")) +
    ggtitle(paste("Salary Amounts for Data Jobs in ", latest_year, " - Offshore Employees")) +
    scale_y_continuous(labels = scales::unit_format(prefix ="$ ", unit=""))
```




```
ggsave("offshore-salaries-2022-by-experience.png")
```

```
## Saving 7 x 5 in image
```