R_Project_DSE5002

Isabella Parlato

2024-11-18

#Objective

Your CEO has decided that the company needs a full-time data scientist, and possibly a team of them in the future. She thinks she needs someone who can help drive data science within then entire organization and could potentially lead a team in the future. She understands that data scientist salaries vary widely across the world and is unsure what to pay them. To complicate matters, salaries are going up due to the great recession and the market is highly competitive. Your CEO has asked you to prepare an analysis on data science salaries and provide them with a range to be competitive and get top talent. The position can work offshore, but the CEO would like to know what the difference is for a person working in the United States. Your company is currently a small company but is expanding rapidly.

Prepare your analysis in an R file. Your final product should be a power point presentation giving your recommendation to the CEO. CEOs do not care about your code and don't want to see it. They want to see visuals and a well thought out analysis. You will need to turn in the power point and the code as a flat R file.

#Restating the questions What is the competitive global pay range for a Full-Time Data Scientist position? What is pay difference for fully offshore remote worker vs a US only employee (keeping in mind that a fully remote person can work offshore)?

#Add data into r

```r
library(readr)
raw_ds_salaries <- read_csv("data/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‘
```

```r
head(raw_ds_salaries)
```

```
## # 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>
```

initial cleaning

```r
#add packages

library(stringr)
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(tidyr)
library(ggplot2)
```

```r
#want to remove columns for "salary" and "salary_currency" & then need to narrow down to only DS positi

ft_ds_only_salaries <- raw_ds_salaries %>%
  select(-salary, -salary_currency) %>%
  filter(job_title == "Data Scientist") %>%
  filter(employment_type == "FT") %>%
  select(-employment_type, -job_title, -"...1")

head(ft_ds_only_salaries)
```

```
## # A tibble: 6 x 7
##   work_year experience_level salary_in_usd employee_residence remote_ratio
##       <dbl> <chr>                    <dbl> <chr>                     <dbl>
## 1      2020 MI                       79833 DE                            0
## 2      2020 MI                       35735 HU                           50
## 3      2020 EN                       51321 FR                            0
## 4      2020 MI                       40481 IN                            0
## 5      2020 EN                       39916 FR                            0
## 6      2020 SE                       68428 GR                          100
## # i 2 more variables: company_location <chr>, company_size <chr>
```

```r
colnames(ft_ds_only_salaries) #used this to figure out name of first column in raw data since it was ju
```

```
## [1] "work_year"          "experience_level"   "salary_in_usd"
## [4] "employee_residence" "remote_ratio"       "company_location"
## [7] "company_size"
```

```r
# now i want to clean up how some of the character variables names

ft_ds_only_salaries <- ft_ds_only_salaries %>%
  mutate(company_size = recode(company_size, "L" = "Large", "M" = "Medium", "S" = "Small"))

ft_ds_only_salaries <- ft_ds_only_salaries %>% mutate(experience_level = recode( experience_level, "EN"
```

```r
#trying to fix order in which positions are pulling
unique(ft_ds_only_salaries$experience_level)
```

```
## [1] "Junior Mid-level"           "Entry-level"
## [3] "Intermediate Senior-level"
```

```r
ft_ds_only_salaries$experience_level <- factor(ft_ds_only_salaries$experience_level, level = c("Entry-le
```

```r
summary_ft_ds_only <- summary(ft_ds_only_salaries)
print(summary_ft_ds_only)
```

```
##    work_year                     experience_level salary_in_usd
##  Min.   :2020   Entry-level              :20      Min.   :  2859
##  1st Qu.:2021   Junior Mid-level         :59      1st Qu.: 55490
##  Median :2022   Intermediate Senior-level:61      Median :104796
##  Mean   :2021                                     Mean   :108923
##  3rd Qu.:2022                                     3rd Qu.:141975
##  Max.   :2022                                     Max.   :412000
##  employee_residence  remote_ratio    company_location   company_size
##  Length:140         Min.   :  0.00   Length:140         Length:140
##  Class :character   1st Qu.:  0.00   Class :character   Class :character
##  Mode  :character   Median :100.00   Mode  :character   Mode  :character
##                     Mean   : 63.93
##                     3rd Qu.:100.00
##                     Max.   :100.00
```

Based on summary, regarding salary_in_usd for only Data Science Roles, regardless of other factors:

1st Qu.: $55,490.00 this is the median of the lower half 3rd Qu.: $141,975.00 this is the median of the upper half Median : $104,796.00 Mean : $108,923.00 Min. : $2,859.00 this feels like a pretty extreme outlier to have Max. : $412,000.00 similar to the min., this is also a significant outlier

```r
#IQR of ft_ds_only_salaries based on summary data

iqr_ft_ds_only_salaries <- 141975.00 - 55490.00

# $86,485.00 interquartile range
```

```r
#next i want to try separating out by company size

ft_ds_small_companies <- ft_ds_only_salaries %>%
  filter(company_size == "Small")
head(ft_ds_small_companies)
```

```
## # A tibble: 6 x 7
##   work_year experience_level salary_in_usd employee_residence remote_ratio
##       <dbl> <fct>                    <dbl> <chr>                     <dbl>
## 1      2020 Entry-level              51321 FR                            0
## 2      2020 Junior Mid-level         45760 PH                          100
## 3      2020 Junior Mid-level         76958 GB                          100
## 4      2020 Entry-level              62726 DE                           50
## 5      2020 Entry-level              49268 DE                            0
## 6      2020 Entry-level             105000 US                          100
## # i 2 more variables: company_location <chr>, company_size <chr>
```

```r
ft_ds_medium_companies <- ft_ds_only_salaries %>%
  filter(company_size == "Medium")
head(ft_ds_medium_companies)
```

```
## # A tibble: 6 x 7
##   work_year experience_level salary_in_usd employee_residence remote_ratio
##       <dbl> <fct>                    <dbl> <chr>                     <dbl>
## 1      2020 Entry-level              39916 FR                            0
## 2      2020 Junior Mid-level         38776 ES                          100
## 3      2020 Junior Mid-level        118000 US                          100
## 4      2020 Junior Mid-level        138350 US                          100
## 5      2021 Entry-level              49646 FR                           50
## 6      2021 Entry-level              80000 US                          100
## # i 2 more variables: company_location <chr>, company_size <chr>
```

```r
ft_ds_large_companies <- ft_ds_only_salaries %>%
  filter(company_size == "Large")
head(ft_ds_large_companies)
```

```
## # A tibble: 6 x 7
##   work_year experience_level       salary_in_usd employee_residence remote_ratio
##       <dbl> <fct>                          <dbl> <chr>                     <dbl>
## 1      2020 Junior Mid-level               79833 DE                            0
## 2      2020 Junior Mid-level               35735 HU                           50
## 3      2020 Junior Mid-level               40481 IN                            0
## 4      2020 Intermediate Senior-l~         68428 GR                          100
## 5      2020 Junior Mid-level              105000 US                          100
## 6      2020 Intermediate Senior-l~        120000 US                           50
## # i 2 more variables: company_location <chr>, company_size <chr>
```

```r
#summaries & IQR by company size for ds_only_salaries

summary(ft_ds_small_companies)
```

```
##    work_year                     experience_level salary_in_usd
##  Min.   :2020   Entry-level              : 6      Min.   :  2859
##  1st Qu.:2020   Junior Mid-level         :12      1st Qu.: 23375
##  Median :2021   Intermediate Senior-level: 2      Median : 50295
##  Mean   :2021                                     Mean   : 53439
##  3rd Qu.:2021                                     3rd Qu.: 83810
##  Max.   :2022                                     Max.   :105000
##  employee_residence  remote_ratio   company_location   company_size
##  Length:20          Min.   :  0.0   Length:20          Length:20
##  Class :character   1st Qu.:  0.0   Class :character   Class :character
##  Mode  :character   Median : 75.0   Mode  :character   Mode  :character
##                     Mean   : 57.5
##                     3rd Qu.:100.0
##                     Max.   :100.0
```

```
iqr_ft_ds_small_companies_salaries <- 83810 - 23375
print(iqr_ft_ds_small_companies_salaries)
```

```
## [1] 60435
```

```
# $60,435.00 IQR for small companies

summary(ft_ds_medium_companies)
```

```
##    work_year                     experience_level salary_in_usd
##  Min.   :2020   Entry-level              : 8      Min.   :  4000
##  1st Qu.:2022   Junior Mid-level         :24      1st Qu.: 88352
##  Median :2022   Intermediate Senior-level:43      Median :130000
##  Mean   :2022                                     Mean   :127084
##  3rd Qu.:2022                                     3rd Qu.:165110
##  Max.   :2022                                     Max.   :260000
##  employee_residence  remote_ratio    company_location   company_size
##  Length:75          Min.   :  0.00   Length:75          Length:75
##  Class :character   1st Qu.:  0.00   Class :character   Class :character
##  Mode  :character   Median :100.00   Mode  :character   Mode  :character
##                     Mean   : 69.33
##                     3rd Qu.:100.00
##                     Max.   :100.00
```

```
iqr_ft_ds_medium_companies_salaries <- 165110 - 88352
print(iqr_ft_ds_medium_companies_salaries)
```

```
## [1] 76758
```

```
# $76,758.00 IQR for medium companies

summary(ft_ds_large_companies)
```

```
##    work_year                     experience_level salary_in_usd
##  Min.   :2020   Entry-level              : 6      Min.   : 13400
##  1st Qu.:2021   Junior Mid-level         :23      1st Qu.: 50000
##  Median :2021   Intermediate Senior-level:16      Median : 90734  
##  Mean   :2021                                     Mean   :103313  
##  3rd Qu.:2022                                     3rd Qu.:135000  
##  Max.   :2022                                     Max.   :412000  
##  employee_residence  remote_ratio    company_location   company_size      
##  Length:45          Min.   :  0.00   Length:45          Length:45         
##  Class :character   1st Qu.: 50.00   Class :character   Class :character  
##  Mode  :character   Median : 50.00   Mode  :character   Mode  :character  
##                     Mean   : 57.78                                        
##                     3rd Qu.:100.00                                        
##                     Max.   :100.00
```

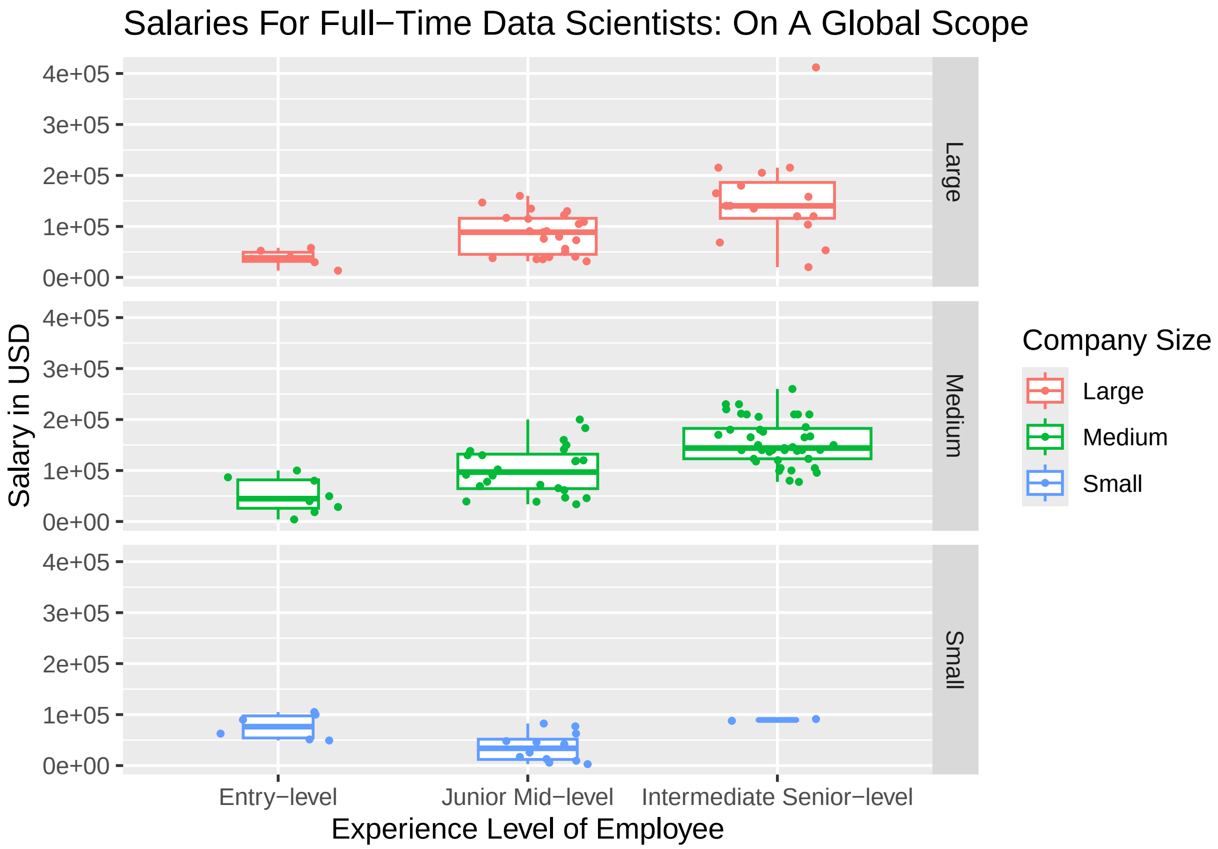
```
iqr_ft_ds_large_companies_salaries <- 135000 - 50000
print(iqr_ft_ds_large_companies_salaries)
```

```
## [1] 85000
```

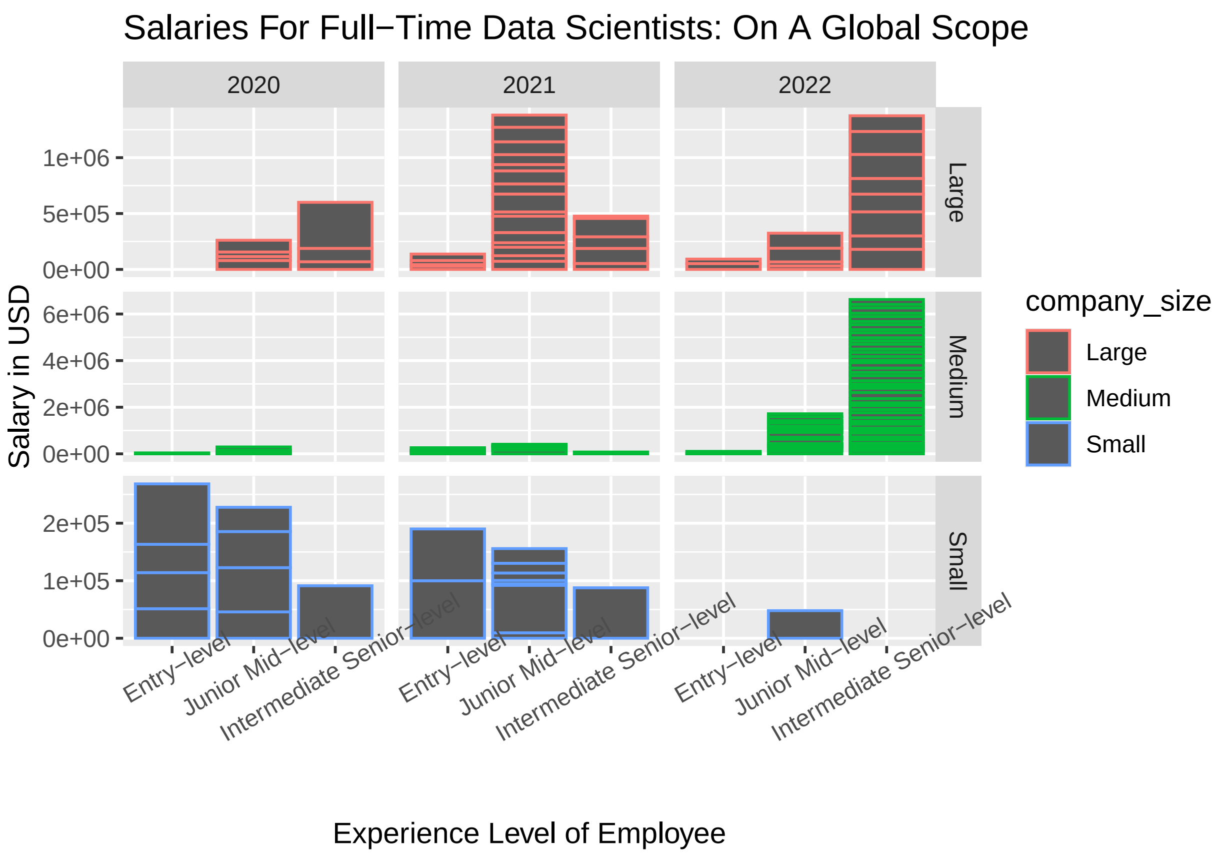
```
# $85,000.00 IQR for large companies

#need to keep in mind that this looks at salary for company size regardless of year and also regardless
```

next I want to try to create a plot to show a visualization of salary range divided by company size and showing separate years and a breakdown by experience level

```
#trying to see how it looks with turkeyplot

ft_ds_only_salaries %>%
  ggplot(mapping = aes(x = experience_level, y = salary_in_usd, color = company_size)) + geom_boxplot(ou
  geom_jitter(position = position_jitterdodge (0.5), size = 0.75) +
  facet_grid(company_size~.) +
  scale_y_continuous(labels = function(x) format(x, big.mark = ",")) +
  labs(x='Experience Level of Employee'
      ,y='Salary in USD'
      ,color = 'Company Size'
      ,title='Salaries For Full-Time Data Scientists: On A Global Scope')
```

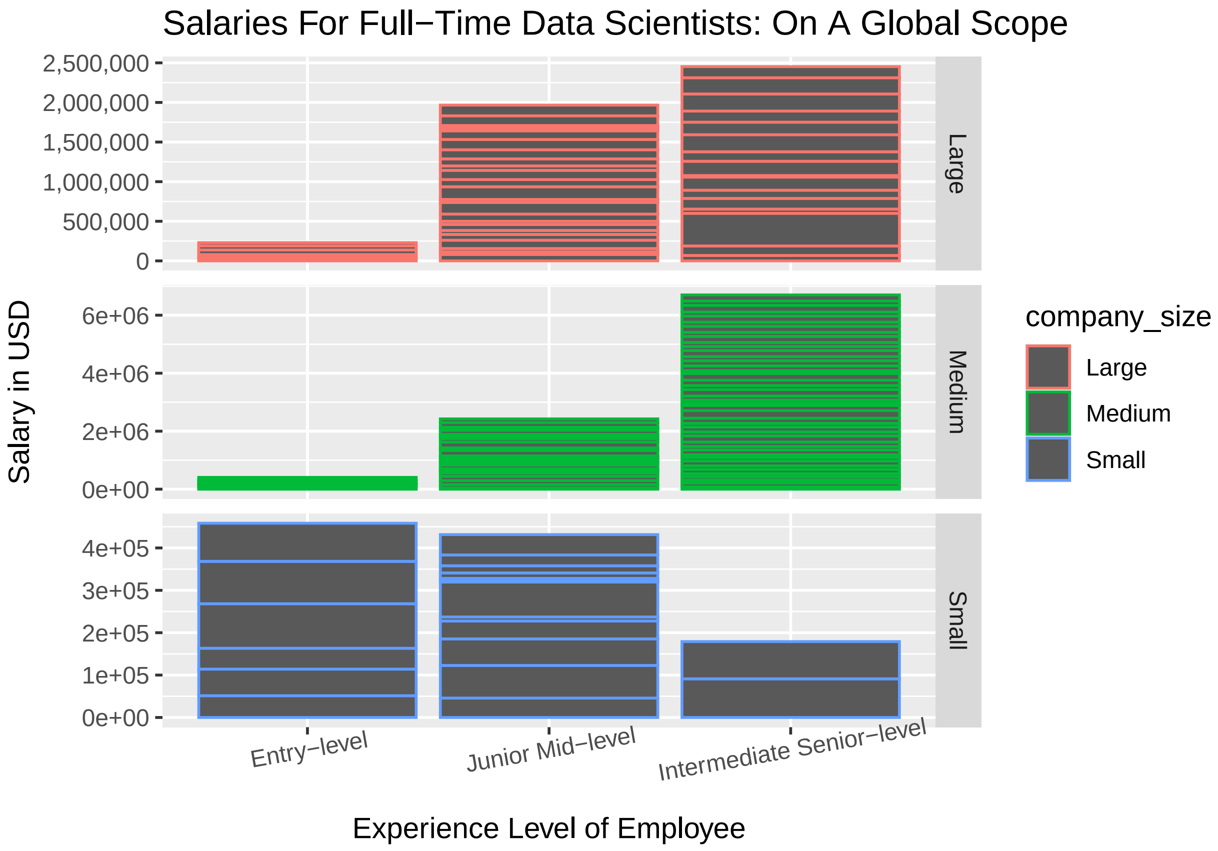


```
# want to see how it looks as a barchart as well with year factored in

ggplot(ft_ds_only_salaries) +
  geom_col(mapping= aes(x = experience_level, y = salary_in_usd, color = company_size)) +   facet_grid(
  theme(axis.text.x = element_text(angle = 30)) +
  scale_y_continuous(labels = function(x) format(x, big.mark = ",")) +
  labs(x='Experience Level of Employee'
      ,y='Salary in USD'
      ,title='Salaries For Full-Time Data Scientists: On A Global Scope')
```



with how the above chart came out, it seems like there isn't much data in some years vs others so we're likely better off not separating by year with the idea that salaries will increase over time

```
# trying without year factored in

ggplot(ft_ds_only_salaries) +
  geom_col(mapping= aes(x = experience_level, y = salary_in_usd, color = company_size)) +  facet_grid(c
  scale_y_continuous(labels = function(x) format(x, big.mark = ",")) +
  theme(axis.text.x = element_text(angle = 10)) +
  labs(x='Experience Level of Employee'
      ,y='Salary in USD'
      ,title='Salaries For Full-Time Data Scientists: On A Global Scope')
```



```
#after comparing all three plots, I think turkeyplot seems like least skewed view - use that in slide a
```

```
#now i need to narrow the DS only positions table down from all global ft to us based workers and full

us_only_ft_ds_salaries <- ft_ds_only_salaries %>%
  filter(employee_residence == "US")
head(us_only_ft_ds_salaries)
```

```
## # A tibble: 6 x 7
##   work_year experience_level       salary_in_usd employee_residence remote_ratio
##       <dbl> <fct>                          <dbl> <chr>                     <dbl>
## 1      2020 Junior Mid-level              105000 US                          100
## 2      2020 Junior Mid-level              118000 US                          100
## 3      2020 Intermediate Senior-l~        120000 US                           50
## 4      2020 Junior Mid-level              138350 US                          100
## 5      2020 Intermediate Senior-l~        412000 US                          100
## 6      2020 Entry-level                   105000 US                          100
## # i 2 more variables: company_location <chr>, company_size <chr>
```

```
summary(us_only_ft_ds_salaries)
```

```
##    work_year                     experience_level salary_in_usd
##  Min.   :2020   Entry-level                 : 6      Min.   : 58000
##  1st Qu.:2021   Junior Mid-level            :21      1st Qu.:120000
##  Median :2022   Intermediate Senior-level:51      Median :140000
##  Mean   :2022                                     Mean   :149408  
##  3rd Qu.:2022                                     3rd Qu.:174500  
##  Max.   :2022                                     Max.   :412000  
##  employee_residence  remote_ratio    company_location   company_size      
##  Length:78          Min.   :  0.00   Length:78          Length:78         
##  Class :character   1st Qu.: 50.00   Class :character   Class :character  
##  Mode  :character   Median :100.00   Mode  :character   Mode  :character  
##                     Mean   : 71.79                                        
##                     3rd Qu.:100.00                                        
##                     Max.   :100.00                                        
```

```
# median: $140,000 - mean: $149,408 - 3rd Q: $174,500 - 1st Q: $120,000 - iqr: $54,500
```

```
non_us_fully_remote_salaries <- ft_ds_only_salaries %>%
  filter(employee_residence != "US") %>%
  filter(remote_ratio == "100")

head(non_us_fully_remote_salaries)
```

```
## # A tibble: 6 x 7
##   work_year experience_level       salary_in_usd employee_residence remote_ratio
##       <dbl> <fct>                          <dbl> <chr>                     <dbl>
## 1      2020 Intermediate Senior-l~         68428 GR                          100
## 2      2020 Junior Mid-level               45760 PH                          100
## 3      2020 Junior Mid-level               76958 GB                          100
## 4      2020 Junior Mid-level               38776 ES                          100
## 5      2021 Junior Mid-level               50000 NG                          100
## 6      2021 Junior Mid-level               75774 CA                          100
## # i 2 more variables: company_location <chr>, company_size <chr>
```

```
summary(non_us_fully_remote_salaries)
```

```
##    work_year                     experience_level salary_in_usd   
##  Min.   :2020   Entry-level              : 5      Min.   :  5679  
##  1st Qu.:2021   Junior Mid-level         :16      1st Qu.: 31615  
##  Median :2021   Intermediate Senior-level: 4      Median : 45760  
##  Mean   :2021                                     Mean   : 51046  
##  3rd Qu.:2022                                     3rd Qu.: 69336  
##  Max.   :2022                                     Max.   :119059  
##  employee_residence  remote_ratio company_location   company_size      
##  Length:25          Min.   :100   Length:25          Length:25         
##  Class :character   1st Qu.:100   Class :character   Class :character  
##  Mode  :character   Median :100   Mode  :character   Mode  :character  
##                     Mean   :100                                        
##                     3rd Qu.:100                                        
##                     Max.   :100                                        
```

```
# median: $45,760 - mean: $51,046 - 3rd Q: $69,336 - 1st Q: $31,615 - iqr: $37,721
```

```
#since the ceo wants to know difference between US workers and that of offshore
```

```r
non_us_salaries_ft_ds <- ft_ds_only_salaries %>%
  filter(employee_residence != "US")

head(non_us_salaries_ft_ds)
```

```
## # A tibble: 6 x 7
##   work_year experience_level       salary_in_usd employee_residence remote_ratio
##       <dbl> <fct>                          <dbl> <chr>                     <dbl>
## 1      2020 Junior Mid-level               79833 DE                            0
## 2      2020 Junior Mid-level               35735 HU                           50
## 3      2020 Entry-level                    51321 FR                            0
## 4      2020 Junior Mid-level               40481 IN                            0
## 5      2020 Entry-level                    39916 FR                            0
## 6      2020 Intermediate Senior-l~         68428 GR                          100
## # i 2 more variables: company_location <chr>, company_size <chr>
```

```r
summary(non_us_salaries_ft_ds)
```

```
##    work_year                     experience_level salary_in_usd
##  Min.   :2020   Entry-level               :14      Min.   :  2859
##  1st Qu.:2021   Junior Mid-level          :38      1st Qu.: 35962
##  Median :2021   Intermediate Senior-level:10      Median : 49823
##  Mean   :2021                                      Mean   : 57989
##  3rd Qu.:2022                                      3rd Qu.: 79296
##  Max.   :2022                                      Max.   :183228
##  employee_residence  remote_ratio    company_location   company_size
##  Length:62          Min.   :  0.00   Length:62          Length:62
##  Class :character   1st Qu.:  0.00   Class :character   Class :character
##  Mode  :character   Median : 50.00   Mode  :character   Mode  :character
##                     Mean   : 54.03
##                     3rd Qu.:100.00
##                     Max.   :100.00
```

```r
# median: $49,823 - mean: $57989 - 3rd Q: $79,296 - 1st Q: $35,962 - iqr: $43,334
```

```r
full_us_offshore_ft_ds_salaries <- us_only_ft_ds_salaries %>%
  full_join(non_us_fully_remote_salaries)
```

```
## Joining with ‘by = join_by(work_year, experience_level, salary_in_usd,
## employee_residence, remote_ratio, company_location, company_size)‘
```

```r
head(full_us_offshore_ft_ds_salaries)
```

```
## # A tibble: 6 x 7
##   work_year experience_level       salary_in_usd employee_residence remote_ratio
##       <dbl> <fct>                          <dbl> <chr>                     <dbl>
## 1      2020 Junior Mid-level              105000 US                          100
## 2      2020 Junior Mid-level              118000 US                          100
## 3      2020 Intermediate Senior-l~        120000 US                           50
## 4      2020 Junior Mid-level              138350 US                          100
## 5      2020 Intermediate Senior-l~        412000 US                          100
## 6      2020 Entry-level                   105000 US                          100
## # i 2 more variables: company_location <chr>, company_size <chr>
```

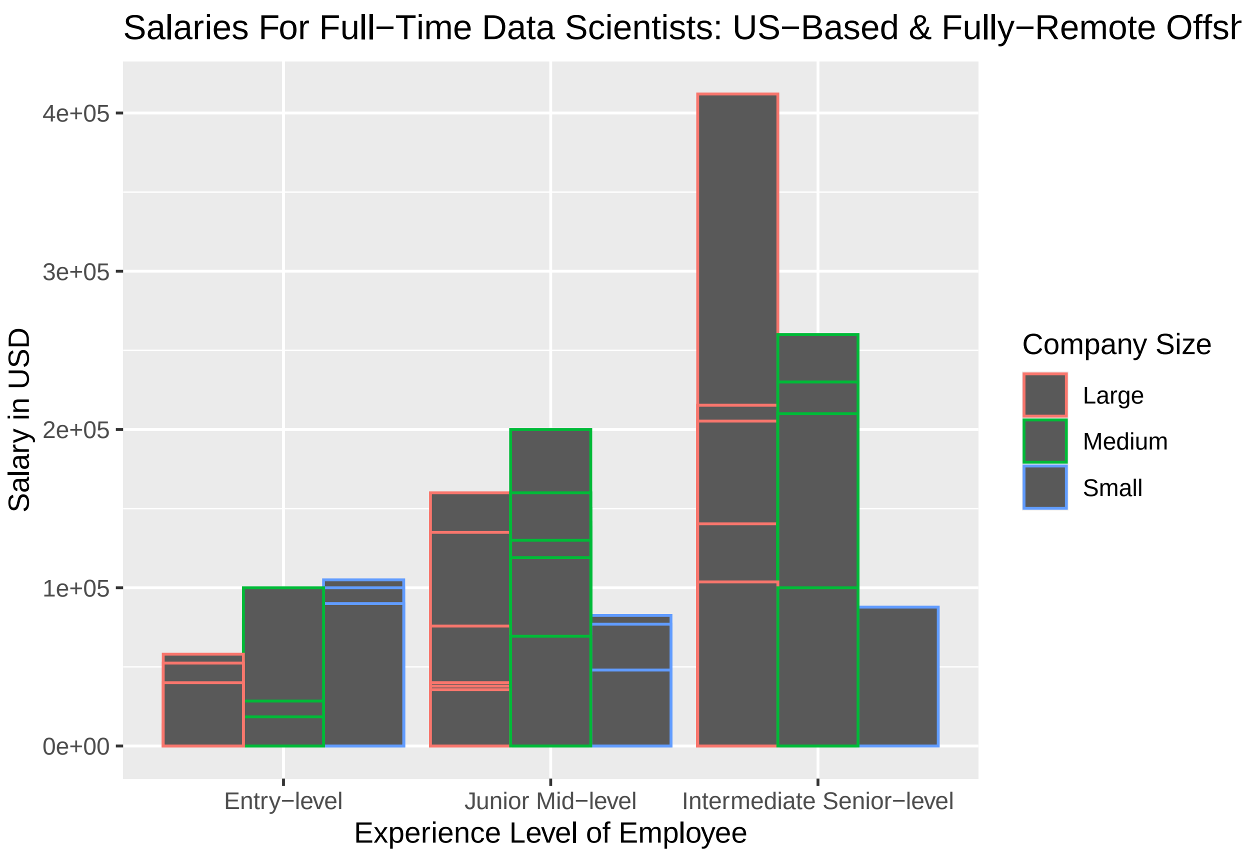
```
summary(full_us_offshore_ft_ds_salaries)
```

```
##    work_year                     experience_level salary_in_usd
##  Min.   :2020   Entry-level              :11      Min.   :  5679
##  1st Qu.:2021   Junior Mid-level         :37      1st Qu.: 81250
##  Median :2022   Intermediate Senior-level:55      Median :130000
##  Mean   :2022                                     Mean   :125534
##  3rd Qu.:2022                                     3rd Qu.:160000
##  Max.   :2022                                     Max.   :412000
##  employee_residence  remote_ratio    company_location   company_size
##  Length:103         Min.   :  0.00   Length:103         Length:103
##  Class :character   1st Qu.:100.00   Class :character   Class :character
##  Mode  :character   Median :100.00   Mode  :character   Mode  :character
##                     Mean   : 78.64
##                     3rd Qu.:100.00
##                     Max.   :100.00
```

salary stats from summary: min - 5,679.00; max - 412,000.00; median - 130,000.00; mean - 125,534.00; 1st Q - 81,250.00; 3rd Q - 160,000.00

```
# IQR for us eligible ds salaries only

iqr_us_offshore_ft_ds_salaries <- 160000.00 - 81250.00
#$78,750
```

```
# now to create a visualization for this

ggplot(full_us_offshore_ft_ds_salaries, mapping= aes(x = experience_level, y = salary_in_usd, color = c
 geom_bar (stat = 'identity', position = 'dodge')   +
  scale_y_continuous(labels = function(x) format(x, big.mark = ",")) +
  labs(x='Experience Level of Employee'
      ,y='Salary in USD'
      ,color = 'Company Size'
      ,title='Salaries For Full-Time Data Scientists: US-Based & Fully-Remote Offshore')
```



```
#now to pull summaries by company size again to get IQR since it seems like there are some outliers inv

large_us_remote_offshore_summary <- full_us_offshore_ft_ds_salaries %>%
  filter (company_size == "Large")
summary(large_us_remote_offshore_summary)
```

```
##    work_year                     experience_level salary_in_usd
##  Min.   :2020   Entry-level              : 4      Min.   : 13400
##  1st Qu.:2021   Junior Mid-level         :14      1st Qu.: 56599
##  Median :2021   Intermediate Senior-level:14      Median :117500
##  Mean   :2021                                     Mean   :118364
##  3rd Qu.:2022                                     3rd Qu.:149800
##  Max.   :2022                                     Max.   :412000
##  employee_residence  remote_ratio    company_location   company_size
##  Length:32          Min.   :  0.00   Length:32          Length:32
##  Class :character   1st Qu.: 37.50   Class :character   Class :character
##  Mode  :character   Median :100.00   Mode  :character   Mode  :character
##                     Mean   : 65.62
##                     3rd Qu.:100.00
##                     Max.   :100.00
```

```
iqr_large_us_remote_offshore_summary <- 149800 - 56599
#IQR large companies: $93,201.00


medium_us_remote_offshore_summary <- full_us_offshore_ft_ds_salaries %>%
  filter (company_size == "Medium")
summary(medium_us_remote_offshore_summary)
```

```
##    work_year                     experience_level salary_in_usd
##  Min.   :2020   Entry-level              : 4      Min.   : 18442
##  1st Qu.:2022   Junior Mid-level         :16      1st Qu.:114723
##  Median :2022   Intermediate Senior-level:40      Median :140000
##  Mean   :2022                                     Mean   :140971
##  3rd Qu.:2022                                     3rd Qu.:171500
##  Max.   :2022                                     Max.   :260000
##  employee_residence  remote_ratio    company_location   company_size
##  Length:60          Min.   :  0.00   Length:60          Length:60
##  Class :character   1st Qu.:100.00   Class :character   Class :character
##  Mode  :character   Median :100.00   Mode  :character   Mode  :character
##                     Mean   : 83.33
##                     3rd Qu.:100.00
##                     Max.   :100.00
```

```
iqr_medium_us_remote_offshore_summary <- 171500 - 114723
#IQR medium companies: $56,777.00


small_us_remote_offshore_summary <- full_us_offshore_ft_ds_salaries %>%
  filter (company_size == "Small")
summary(small_us_remote_offshore_summary)
```

```
##    work_year                     experience_level salary_in_usd
##  Min.   :2020   Entry-level              :3       Min.   :  5679
##  1st Qu.:2020   Junior Mid-level         :7       1st Qu.: 35646
##  Median :2021   Intermediate Senior-level:1       Median : 76958
##  Mean   :2021                                     Mean   : 62188
##  3rd Qu.:2021                                     3rd Qu.: 88869
##  Max.   :2022                                     Max.   :105000
##  employee_residence  remote_ratio    company_location   company_size
##  Length:11          Min.   :  0.00   Length:11          Length:11
##  Class :character   1st Qu.:100.00   Class :character   Class :character
##  Mode  :character   Median :100.00   Mode  :character   Mode  :character
##                     Mean   : 90.91
##                     3rd Qu.:100.00
##                     Max.   :100.00
```

```
iqr_small_us_remote_offshore_summary <- 88869 - 35646
#IQR small companies: $53,223.00
```

#looking ahead/expanding beyond for perspective

It feels like looking solely at DS only positions isn't much data so in an effort to help give the CEO perspective about salaries for data roles, I want to pull the 5 most popular job positions plus DS if it isn't one of them from the raw data, once again limit down to FT only since that's the type of role that she wants. I also think I can remove large companies since we're a small company on our way to becoming medium sized.

```
#figure out top 5 positions

raw_ds_salaries %>% count(job_title)
```

```
## # A tibble: 50 x 2
##    job_title                                n
##    <chr>                                <int>
##  1 3D Computer Vision Researcher            1
##  2 AI Scientist                             7
##  3 Analytics Engineer                       4
##  4 Applied Data Scientist                   5
##  5 Applied Machine Learning Scientist       4
##  6 BI Data Analyst                          6
##  7 Big Data Architect                       1
##  8 Big Data Engineer                        8
##  9 Business Data Analyst                    5
## 10 Cloud Data Engineer                      2
## # i 40 more rows
```

```
# Data Analyst - 97; Data Engineer - 132; Data Scientist - 143; Machine Learning Engineer - 41; Researc
```

```
#now to clean up df to just those 5 positions and remove large companies and clean up column variables

top_five_positions_global_no_large_companies <- raw_ds_salaries %>%
  select(-salary, -salary_currency) %>%
  filter(job_title %in% c("Data Scientist", "Data Analyst", "Data Engineer", "Machine Learning Engineer
  filter(employment_type == "FT") %>%
  select(-employment_type, -"...1") %>%
  filter (company_size != "L")

top_five_positions_global_no_large_companies <- top_five_positions_global_no_large_companies %>%
  mutate(company_size = recode(company_size, "M" = "Medium", "S" = "Small"))

top_five_positions_global_no_large_companies <- top_five_positions_global_no_large_companies %>% mutate

head(top_five_positions_global_no_large_companies)
```

```
## # A tibble: 6 x 8
##   work_year experience_level          job_title salary_in_usd employee_residence
##       <dbl> <chr>                     <chr>             <dbl> <chr>
## 1      2020 Entry-level               Data Sci~         51321 FR
## 2      2020 Entry-level               Data Sci~         39916 FR
## 3      2020 Entry-level               Data Eng~         41689 JP
## 4      2020 Junior Mid-level          Machine ~         43331 CN
## 5      2020 Intermediate Senior-level Data Eng~         33511 MX
## 6      2020 Junior Mid-level          Research~        450000 US
## # i 3 more variables: remote_ratio <dbl>, company_location <chr>,
## #   company_size <chr>
```

```
summary(top_five_positions_global_no_large_companies)
```

```
##    work_year    experience_level    job_title          salary_in_usd
##  Min.   :2020   Length:306         Length:306         Min.   :  2859
##  1st Qu.:2021   Class :character   Class :character   1st Qu.: 65949
##  Median :2022   Mode  :character   Mode  :character   Median :102100
##  Mean   :2022                                         Mean   :108702
##  3rd Qu.:2022                                         3rd Qu.:140000
##  Max.   :2022                                         Max.   :450000
##  employee_residence  remote_ratio    company_location   company_size
##  Length:306         Min.   :  0.00   Length:306         Length:306
##  Class :character   1st Qu.:  0.00   Class :character   Class :character
##  Mode  :character   Median :100.00   Mode  :character   Mode  :character
##                     Mean   : 71.41
##                     3rd Qu.:100.00
##                     Max.   :100.00
```

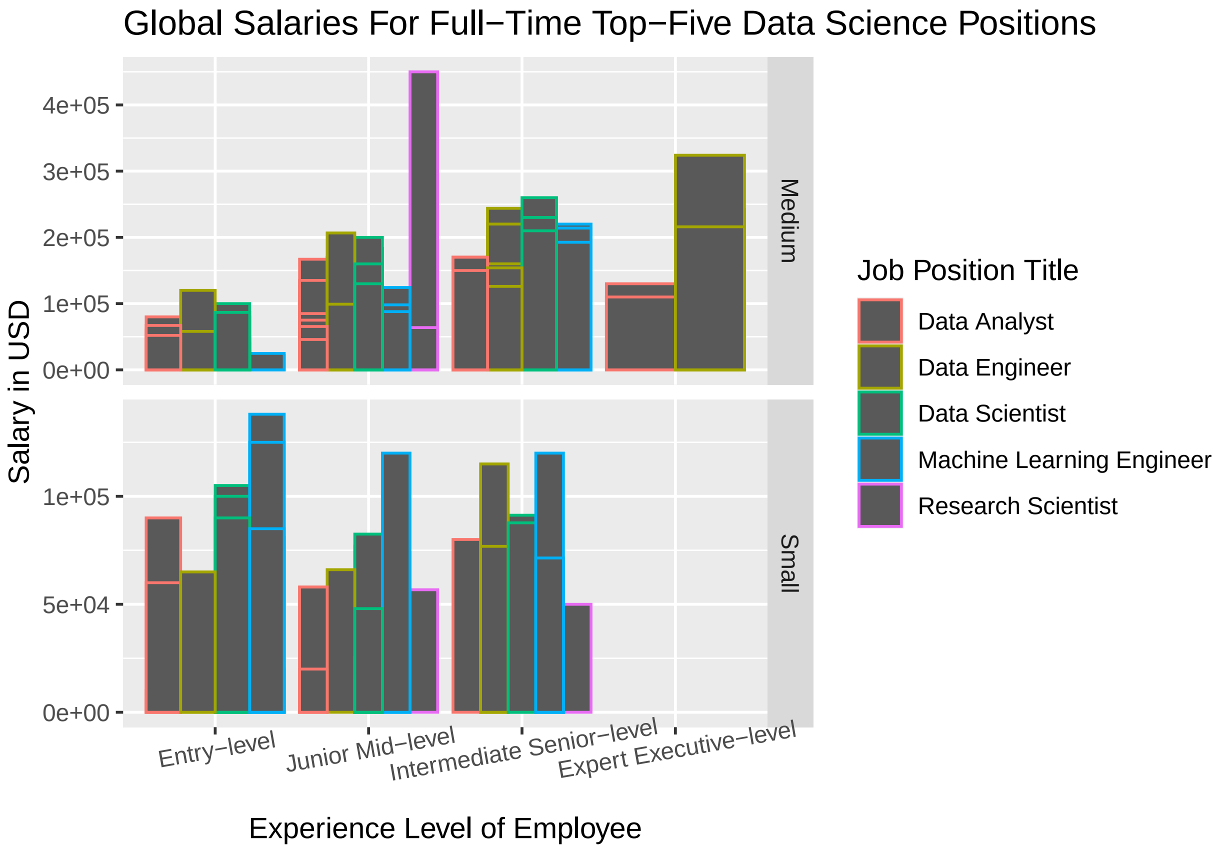
```
#mean: 108702 - median: 102100 - 3rd Q: 140000 - 1st Q: 65949 - IQR: 74051
```

```
#trying to fix order in which positions are pulling so correct on plot

unique(top_five_positions_global_no_large_companies$experience_level)
```

```
## [1] "Entry-level"               "Junior Mid-level"
## [3] "Intermediate Senior-level" "Expert Executive-level"
```

```
top_five_positions_global_no_large_companies$experience_level <- factor(top_five_positions_global_no_la
```

```
#now to create a visualization

ggplot(top_five_positions_global_no_large_companies, mapping= aes(x = experience_level, y = salary_in_us
 geom_bar (stat = 'identity', position = 'dodge')    +
  facet_grid(company_size~., scales = "free_y") +
  scale_y_continuous(labels = function(x) format(x, big.mark = ",")) +
  theme(axis.text.x = element_text(angle = 10)) +
  labs(x='Experience Level of Employee'
      ,y='Salary in USD'
      ,color = "Job Position Title"
      ,title='Global Salaries For Full-Time Top-Five Data Science Positions')
```



```
#now to break it down by company size as well

small_companies_top_five_global <- top_five_positions_global_no_large_companies %>%
  filter (company_size == "Small")
summary(small_companies_top_five_global)
```

```
##    work_year                        experience_level  job_title
##  Min.   :2020   Entry-level              :17      Length:48
##  1st Qu.:2020   Junior Mid-level         :20      Class :character
##  Median :2021   Intermediate Senior-level:11      Mode  :character
##  Mean   :2021   Expert Executive-level   : 0
##  3rd Qu.:2021
##  Max.   :2022
##  salary_in_usd    employee_residence  remote_ratio    company_location
##  Min.   :  2859   Length:48          Min.   :  0.00   Length:48
##  1st Qu.: 42070   Class :character   1st Qu.: 50.00   Class :character
##  Median : 61363   Mode  :character   Median :100.00   Mode  :character
##  Mean   : 61416                      Mean   : 68.75
##  3rd Qu.: 83125                      3rd Qu.:100.00
##  Max.   :138000                      Max.   :100.00
##  company_size
##  Length:48
##  Class :character
##  Mode  :character
##
##
```

```
## 
```

```r
iqr_small_companies_top_five_global <- 83125 - 42070
#median: $61,363 - mean: $61,416 - IQR: $41,055


medium_companies_top_five_global <- top_five_positions_global_no_large_companies %>%
  filter (company_size == "Medium")
summary(medium_companies_top_five_global)
```

```
##    work_year                     experience_level  job_title        
##  Min.   :2020   Entry-level              : 21     Length:258        
##  1st Qu.:2022   Junior Mid-level         : 80     Class :character  
##  Median :2022   Intermediate Senior-level:151     Mode  :character  
##  Mean   :2022   Expert Executive-level   :  6                       
##  3rd Qu.:2022                                                       
##  Max.   :2022                                                       
##  salary_in_usd    employee_residence  remote_ratio   company_location  
##  Min.   :  4000   Length:258         Min.   :  0.0   Length:258        
##  1st Qu.: 78526   Class :character   1st Qu.:  0.0   Class :character  
##  Median :113950   Mode  :character   Median :100.0   Mode  :character  
##  Mean   :117499                      Mean   : 71.9                     
##  3rd Qu.:150000                      3rd Qu.:100.0                     
##  Max.   :450000                      Max.   :100.0                     
##  company_size      
##  Length:258        
##  Class :character  
##  Mode  :character  
##                    
##                    
## 
```

```r
iqr_medium_companies_top_five_global <- 150000 - 78526
#median: $113,950 - mean: $117,499 - IQR: $71,474
```