Project 1

2024-11-07

Project 1 rubric: Original Question from the CEO

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.

Restated Question From CEO

Your CEO wants you to give them a range of salaries for all full time data science positions for this year, which is almost 2025. Their company is small and they are wanting someone that will grow with their company, eventually becoming a lead. They would prefer to have someone that is located in the US, but they are willing to be flexible and would also like the salary range of someone working outside the US.

-Since the CEO is wanting someone that will grow with their company, they are asking for the salaries of Entry to Mid Level positions.

-Since the CEO's company is currently small, they should be looking at salaries of small companies.

-Need to find the percentage of salary increase over the years to calculate what the current salaries would be for 2025.

My Questions:

1). What are the salaries of full time data scientist positions by experience over the years?

2). For the estimated 2025 salary percent increase, what are the Median and Interquartile salary ranges of full time data scientists by experience levels?

3). For the estimated 2025 salary percent increase, what are the Entry and Mid level salaries of full time data scientists in the US vs not in the US among the different company sizes?

4). For the estimated 2025 salary percent increase, what are the Median and Interquartile salary ranges of Entry & Mid level full time data scientists salaries among small companies by location?

Opening file containing salary data:

```
#file.choose()
infile = "/Users/krishabugajski/Desktop/R_Python/Project_1/R_Project_DSE5002/r+project+data.csv"
salaries = read.csv(infile)

head(salaries)
```

```
##   X work_year experience_level employment_type                 job_title
## 1 0      2020               MI              FT             Data Scientist
## 2 1      2020               SE              FT Machine Learning Scientist
## 3 2      2020               SE              FT          Big Data Engineer
## 4 3      2020               MI              FT       Product Data Analyst
## 5 4      2020               SE              FT  Machine Learning Engineer
## 6 5      2020               EN              FT               Data Analyst
##   salary salary_currency salary_in_usd employee_residence remote_ratio
## 1  70000             EUR         79833                 DE            0
## 2 260000             USD        260000                 JP            0
## 3  85000             GBP        109024                 GB           50
## 4  20000             USD         20000                 HN            0
## 5 150000             USD        150000                 US           50
## 6  72000             USD         72000                 US          100
##   company_location company_size
## 1               DE            L
## 2               JP            S
## 3               GB            M
## 4               HN            S
## 5               US            L
## 6               US            L
```

Exploratory Data Analysis:

```
summary(salaries)
```

```
##        X           work_year    experience_level   employment_type
##  Min.   :  0.0   Min.   :2020   Length:607         Length:607
##  1st Qu.:151.5   1st Qu.:2021   Class :character   Class :character
##  Median :303.0   Median :2022   Mode  :character   Mode  :character
##  Mean   :303.0   Mean   :2021
##  3rd Qu.:454.5   3rd Qu.:2022
##  Max.   :606.0   Max.   :2022
##   job_title            salary         salary_currency    salary_in_usd
##  Length:607         Min.   :    4000   Length:607         Min.   :  2859
##  Class :character   1st Qu.:   70000   Class :character   1st Qu.: 62726
##  Mode  :character   Median :  115000   Mode  :character   Median :101570
##                     Mean   :  324000                      Mean   :112298
##                     3rd Qu.:  165000                      3rd Qu.:150000
##                     Max.   :30400000                      Max.   :600000
##  employee_residence  remote_ratio    company_location   company_size
##  Length:607         Min.   :  0.00   Length:607         Length:607
##  Class :character   1st Qu.: 50.00   Class :character   Class :character
##  Mode  :character   Median :100.00   Mode  :character   Mode  :character
##                     Mean   : 70.92
##                     3rd Qu.:100.00
##                     Max.   :100.00
```

```
str(salaries)
```

```
## 'data.frame':    607 obs. of  12 variables:
##  $ X                 : int  0 1 2 3 4 5 6 7 8 9 ...
##  $ work_year         : int  2020 2020 2020 2020 2020 2020 2020 2020 2020 2020 ...
##  $ experience_level  : chr  "MI" "SE" "SE" "MI" ...
##  $ employment_type   : chr  "FT" "FT" "FT" "FT" ...
##  $ job_title         : chr  "Data Scientist" "Machine Learning Scientist" "Big Data E
ngineer" "Product Data Analyst" ...
##  $ salary            : int  70000 260000 85000 20000 150000 72000 190000 11000000 135
000 125000 ...
##  $ salary_currency   : chr  "EUR" "USD" "GBP" "USD" ...
##  $ salary_in_usd     : int  79833 260000 109024 20000 150000 72000 190000 35735 13500
0 125000 ...
##  $ employee_residence: chr  "DE" "JP" "GB" "HN" ...
##  $ remote_ratio      : int  0 0 50 0 50 100 100 50 100 50 ...
##  $ company_location  : chr  "DE" "JP" "GB" "HN" ...
##  $ company_size      : chr  "L" "S" "M" "S" ...
```

Comments on variables

work_year <- numeric values, ranges from 2020 to 2022

experience_level <- contains 4 character values

employment_type <- contains 4 character values

job_title <- Several Job titles as characters

salary <- contains numeric salaries in several currencies

salary_currency <- character values of currency types

salary_in_usd <- contains numeric salaries in US currency

employee_residence <- character values that states employee location

company_location <- character values that state company location

remote_ratio <- numeric values 1/5/100, representing remote/part remote/not remote

company_size <- contains 3 character values

Changing work_year,experience_level, employment_type, remote_ratio, and company_size to be factors

```
salaries$work_year <- factor(salaries$work_year)

salaries$experience_level <- factor(salaries$experience_level)

salaries$employment_type <- factor(salaries$employment_type)

salaries$remote_ratio <- factor(salaries$remote_ratio)

salaries$company_size <- factor(salaries$company_size)

str(salaries)
```

```
## 'data.frame':    607 obs. of  12 variables:
##  $ X                 : int  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","EX","MI",..: 3 4 4 3 4 1 4 3 3 4 ...
##  $ employment_type   : Factor w/ 4 levels "CT","FL","FT",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ job_title         : chr  "Data Scientist" "Machine Learning Scientist" "Big Data E
ngineer" "Product Data Analyst" ...
##  $ salary            : int  70000 260000 85000 20000 150000 72000 190000 11000000 135
000 125000 ...
##  $ salary_currency   : chr  "EUR" "USD" "GBP" "USD" ...
##  $ salary_in_usd     : int  79833 260000 109024 20000 150000 72000 190000 35735 13500
0 125000 ...
##  $ employee_residence: chr  "DE" "JP" "GB" "HN" ...
##  $ remote_ratio      : Factor w/ 3 levels "0","50","100": 1 1 2 1 2 3 3 2 3 2 ...
##  $ company_location  : chr  "DE" "JP" "GB" "HN" ...
##  $ company_size      : Factor w/ 3 levels "L","M","S": 1 3 2 3 1 1 3 1 1 3 ...
```

Packages we will be using

```
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
```

```
library(ggplot2)
library(scales)
library(knitr)
library(kableExtra)
```

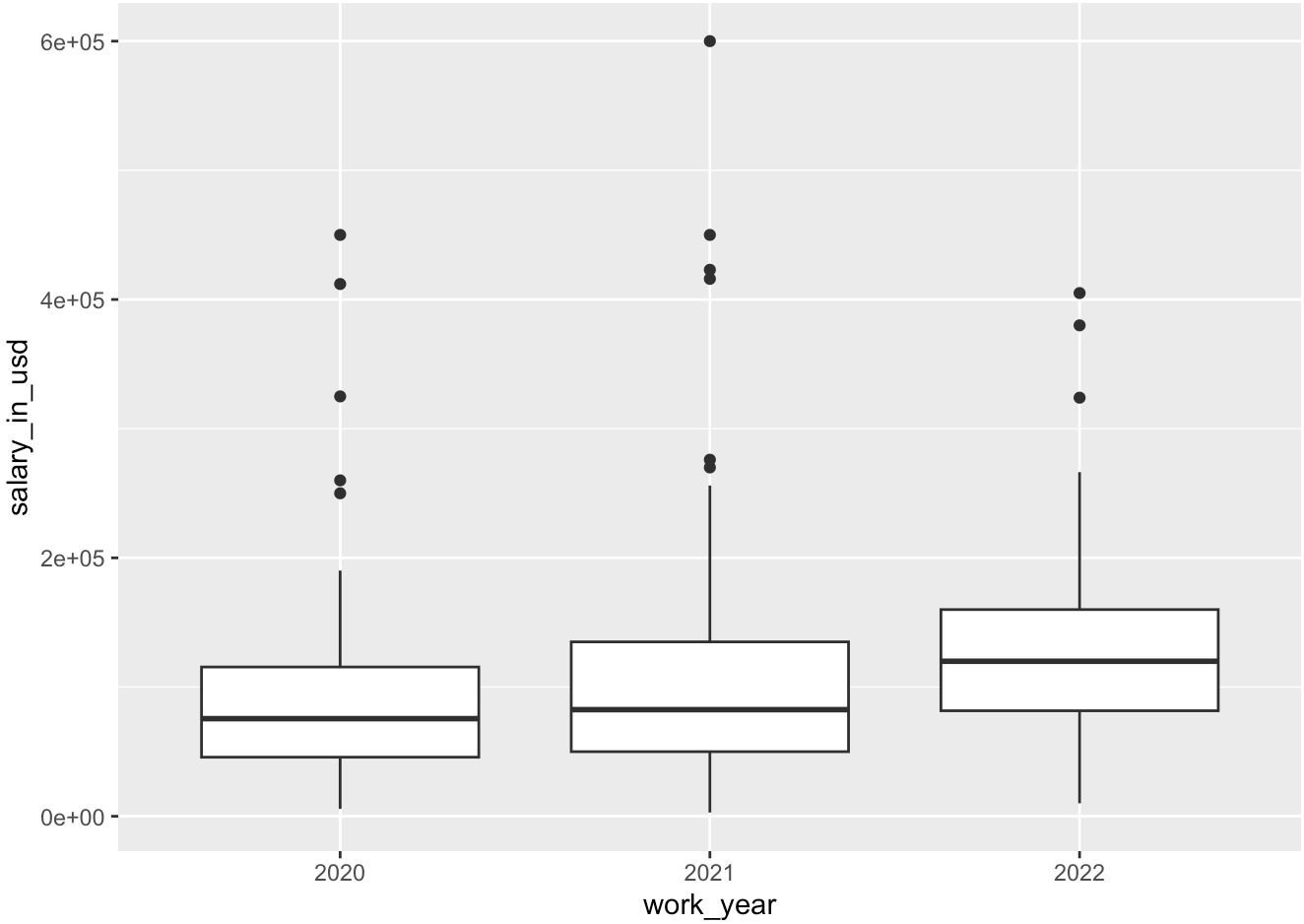
```
##
## Attaching package: 'kableExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     group_rows
```

Plots for the Variables:

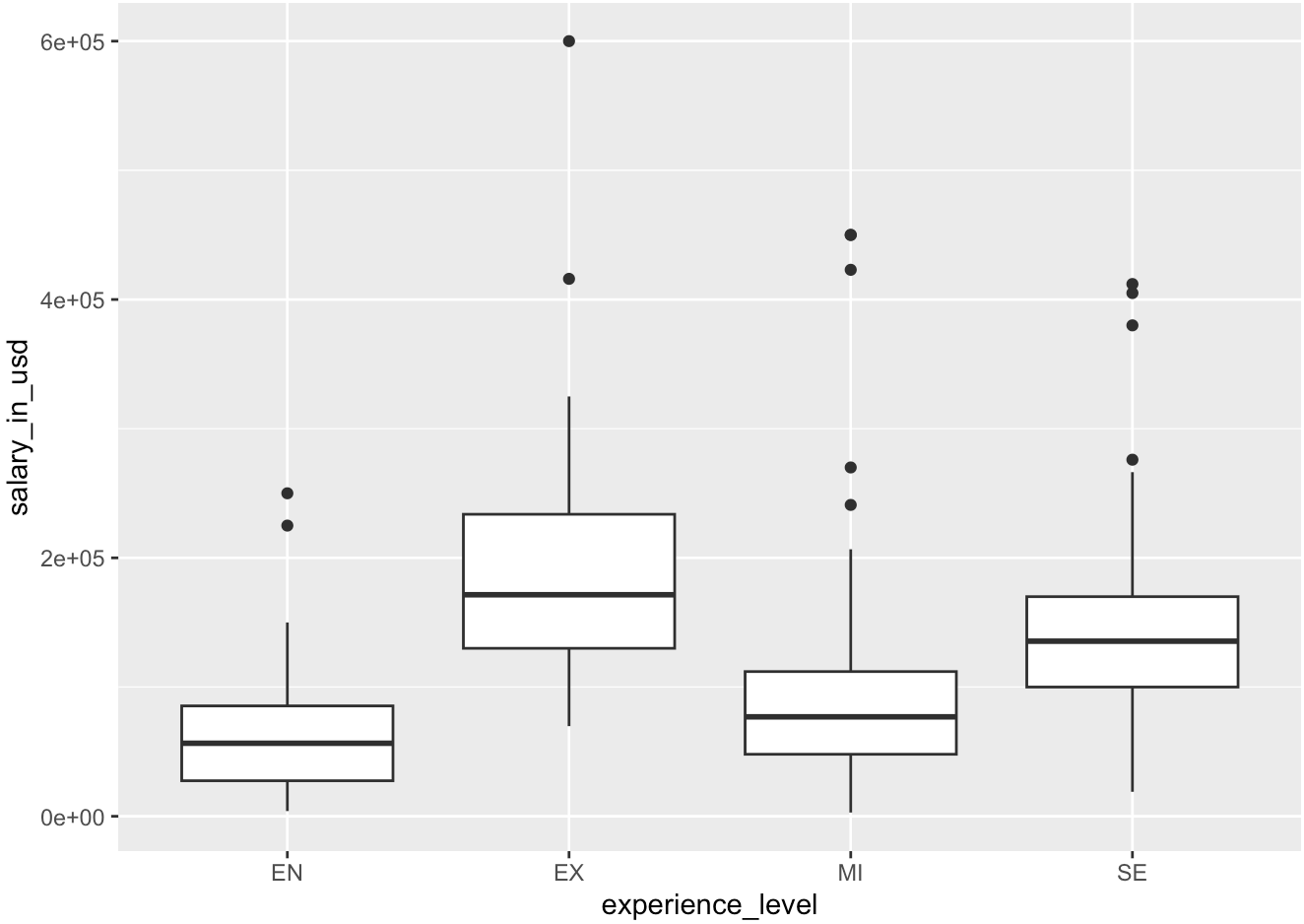
Plot for variable work_year

```
require("ggplot2")
ggplot(salaries, aes(x= work_year, y = salary_in_usd))+
  geom_boxplot()
```



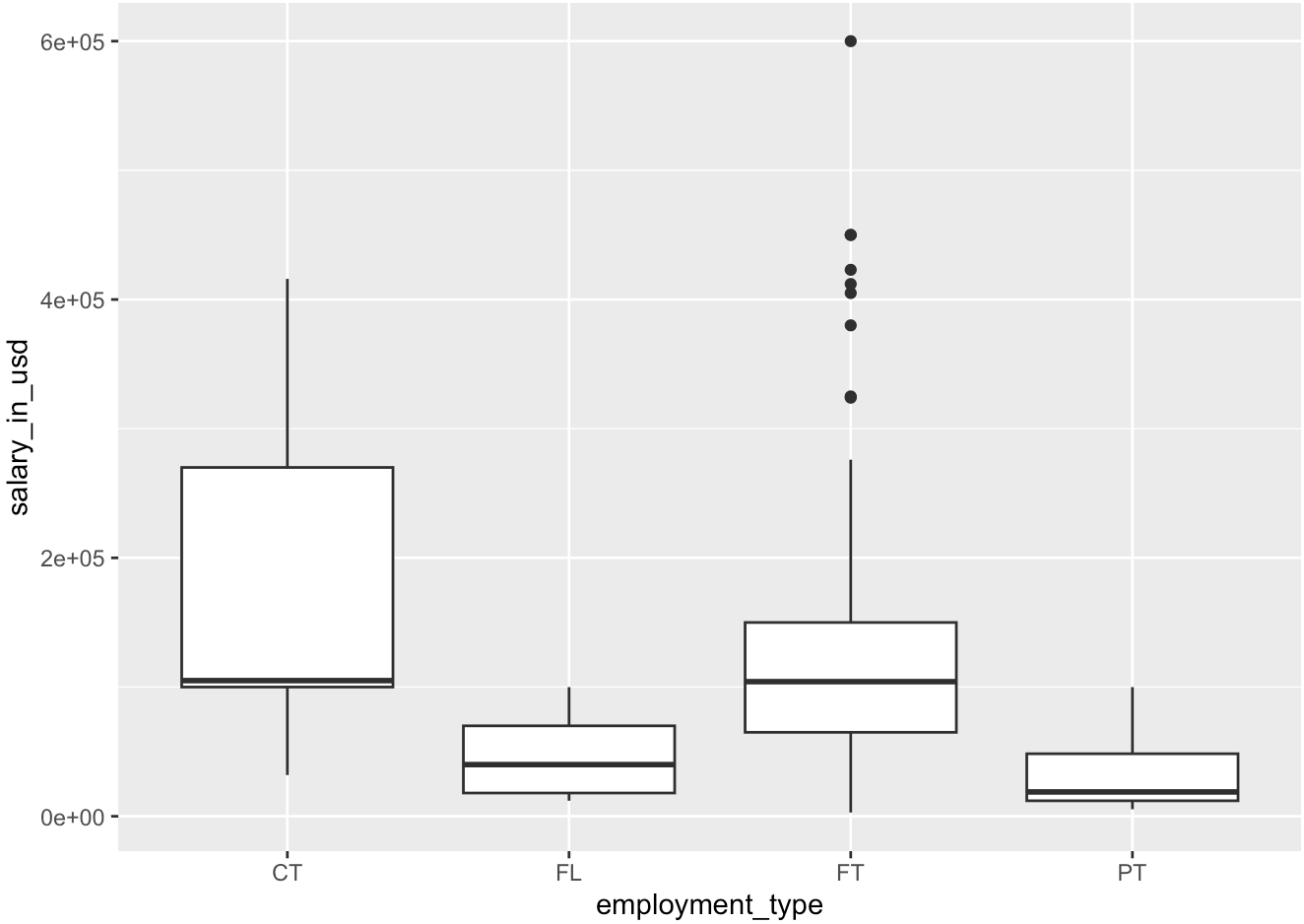
Plot for variable experience_level

```
ggplot(salaries, aes(x= experience_level, y = salary_in_usd))+
  geom_boxplot()
```



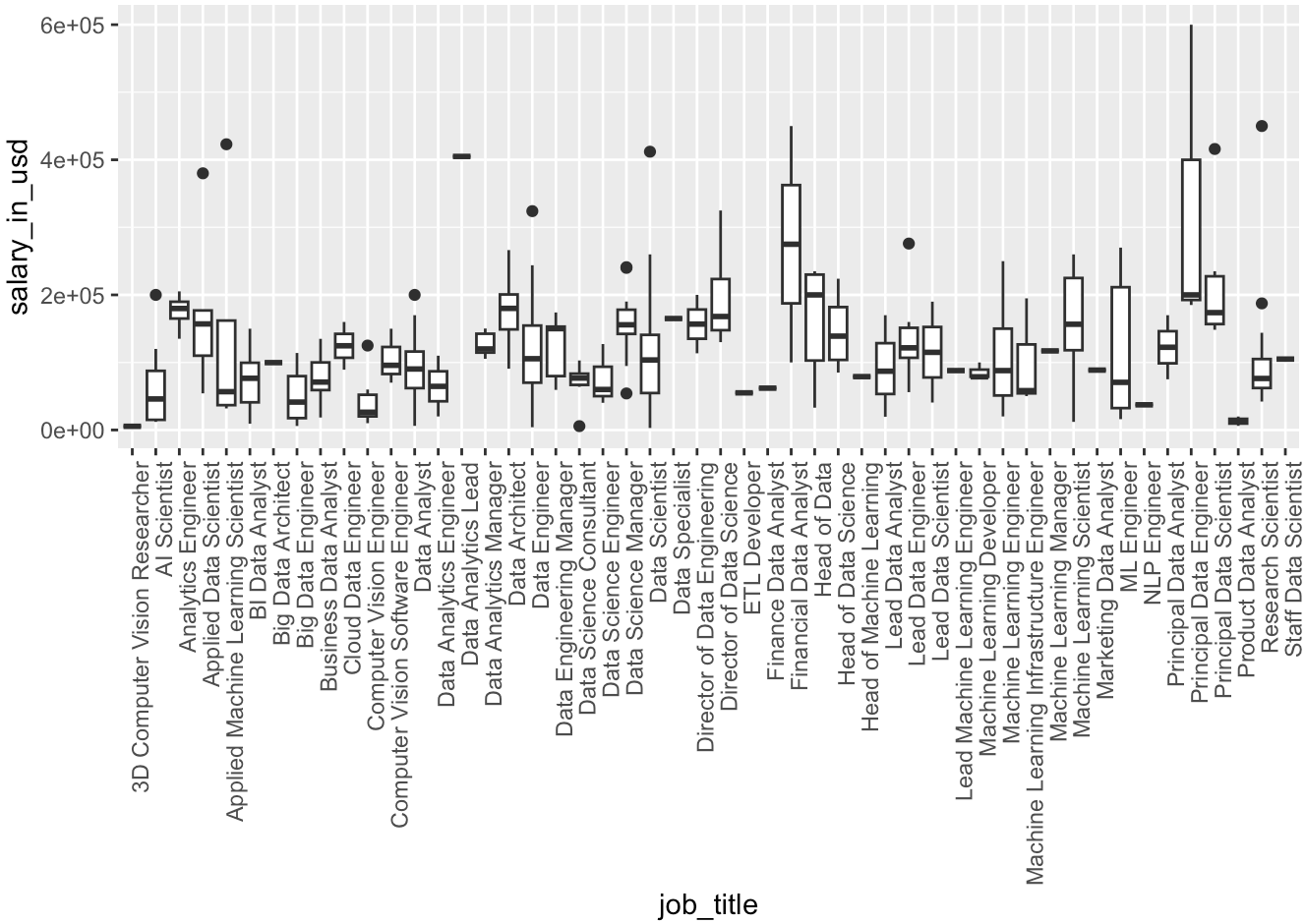
Plot for variable employment_type

```
ggplot(salaries, aes(x= employment_type, y = salary_in_usd))+
  geom_boxplot()
```



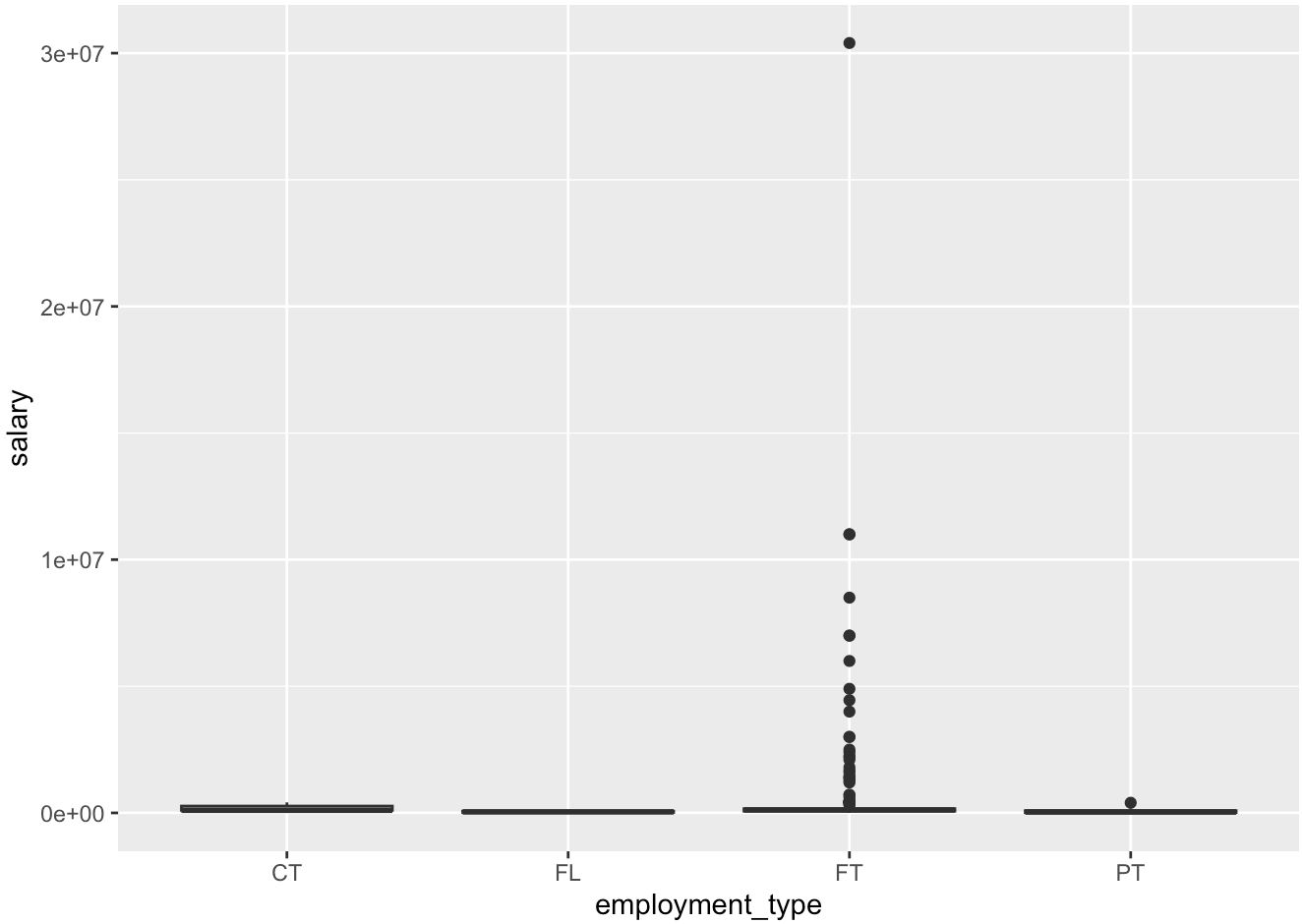
Plot for variable job_title

```
ggplot(salaries, aes(x= job_title, y = salary_in_usd))+
  geom_boxplot()+
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```



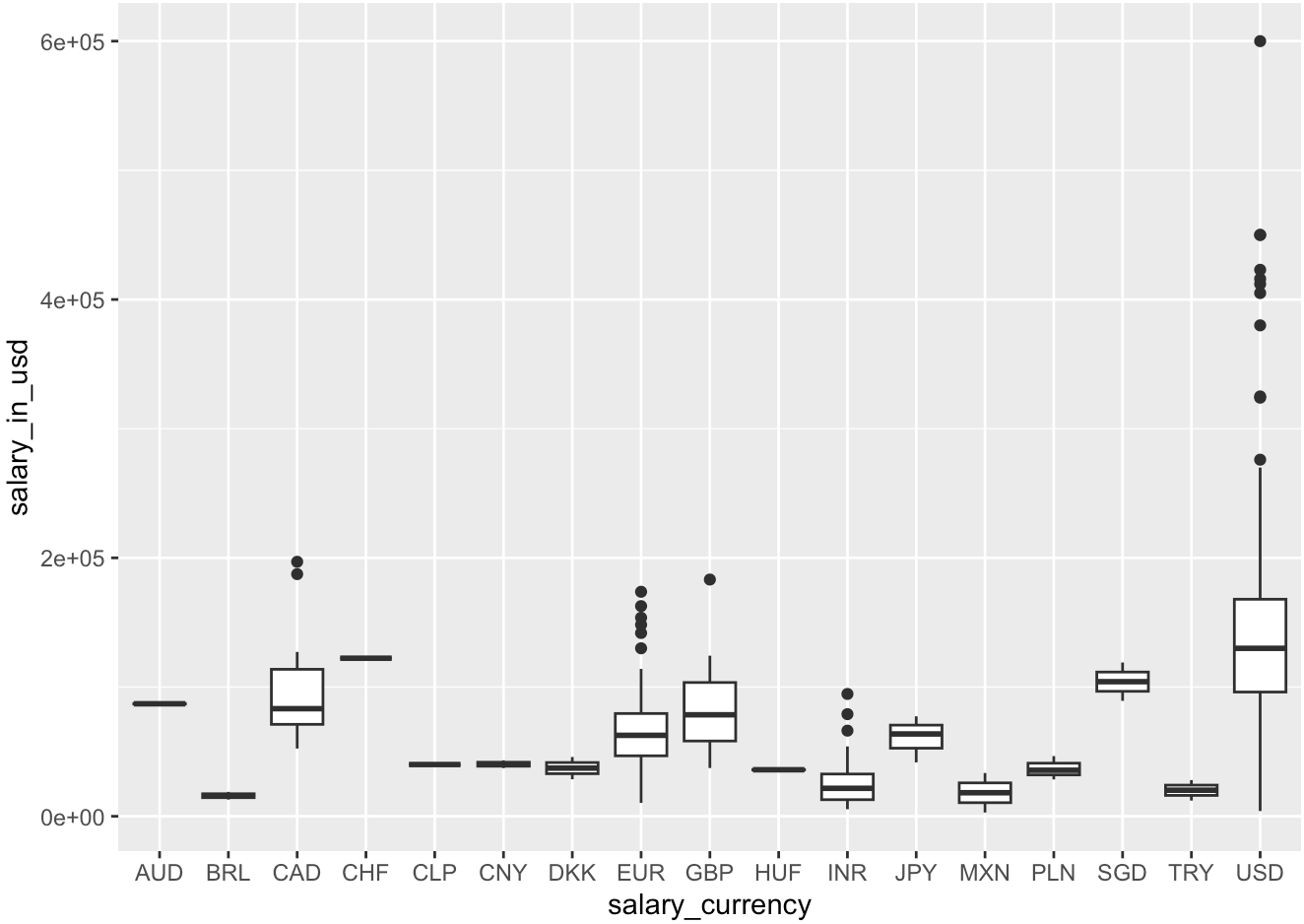
Plot for variable salary

```
ggplot(salaries, aes(x= employment_type, y = salary))+
  geom_boxplot()
```



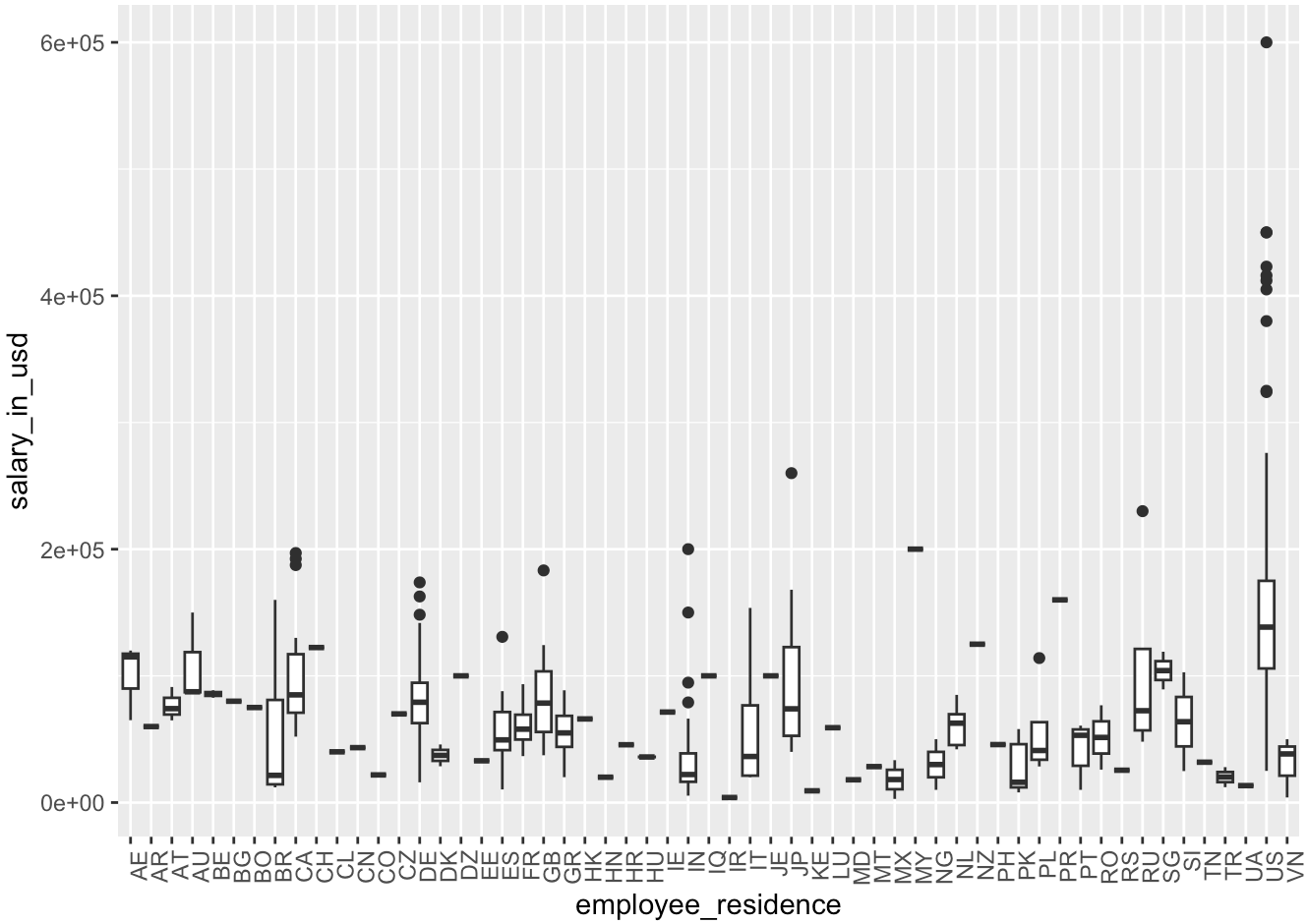
Plot for variable salary_currency

```
ggplot(salaries, aes(x= salary_currency, y = salary_in_usd))+
  geom_boxplot()
```



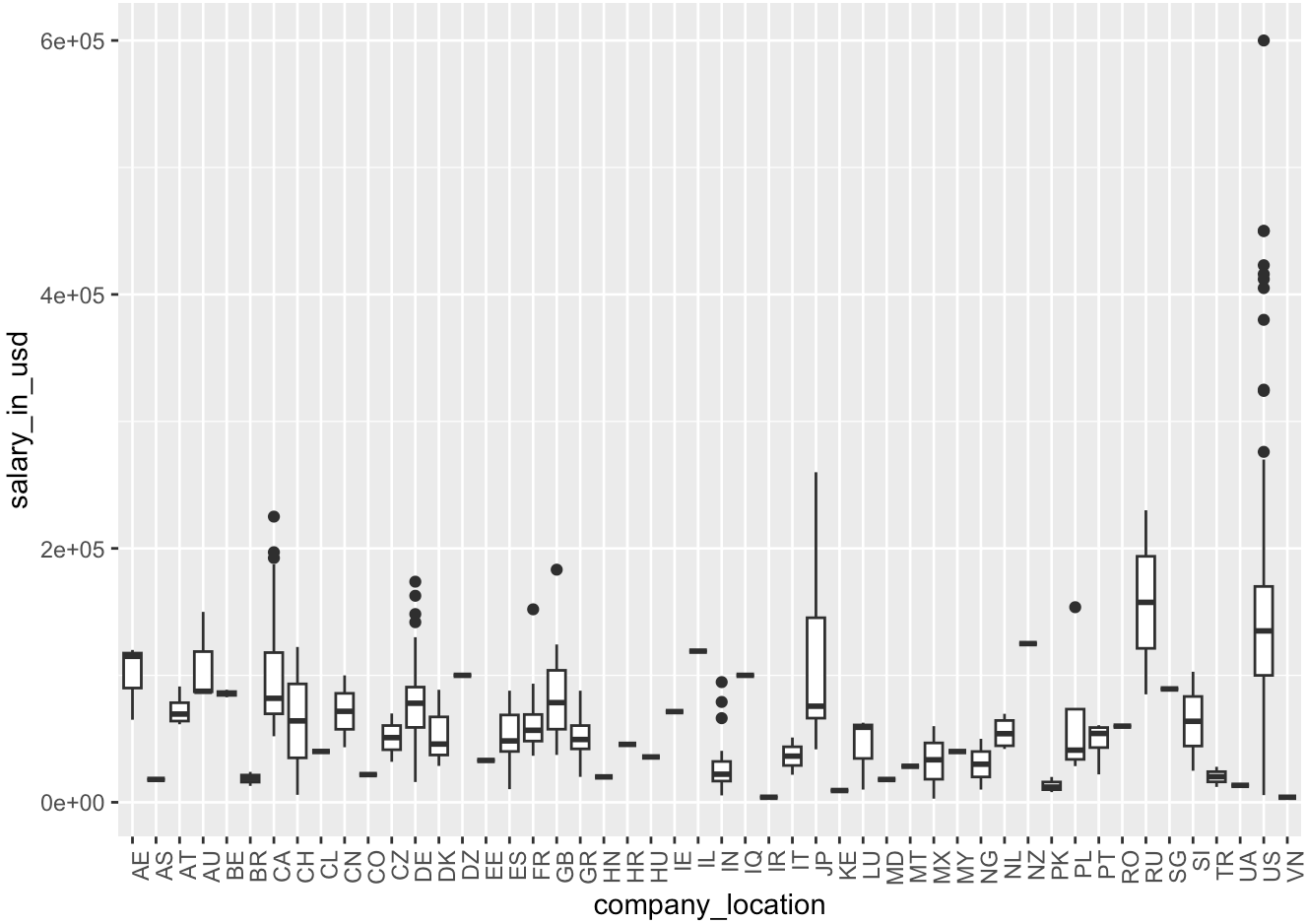
Plot for variable employee_residence

```
ggplot(salaries, aes(x= employee_residence, y = salary_in_usd))+
  geom_boxplot()+
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```



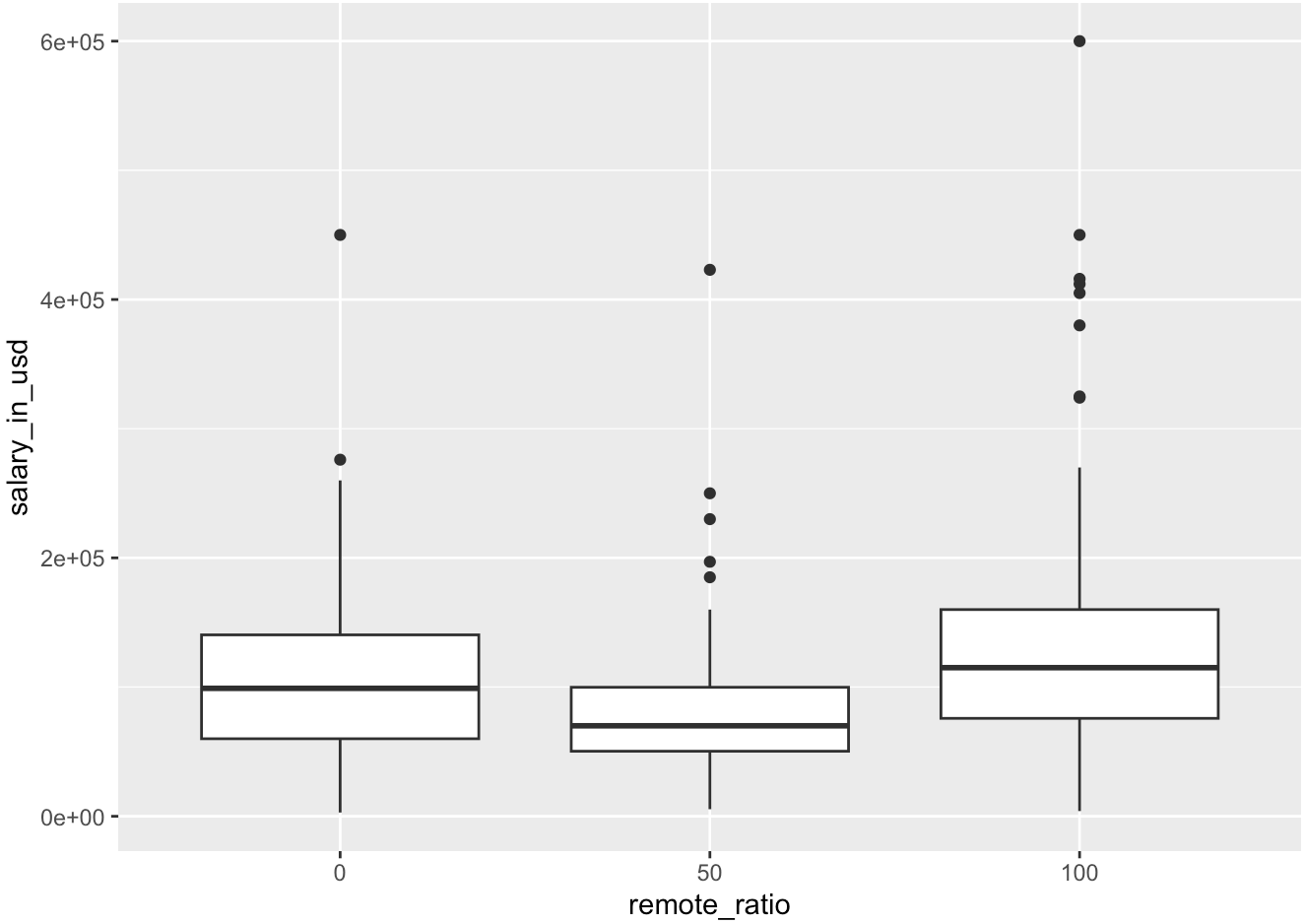
Plot for variable company_location

```
ggplot(salaries, aes(x= company_location, y = salary_in_usd))+
  geom_boxplot()+
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```



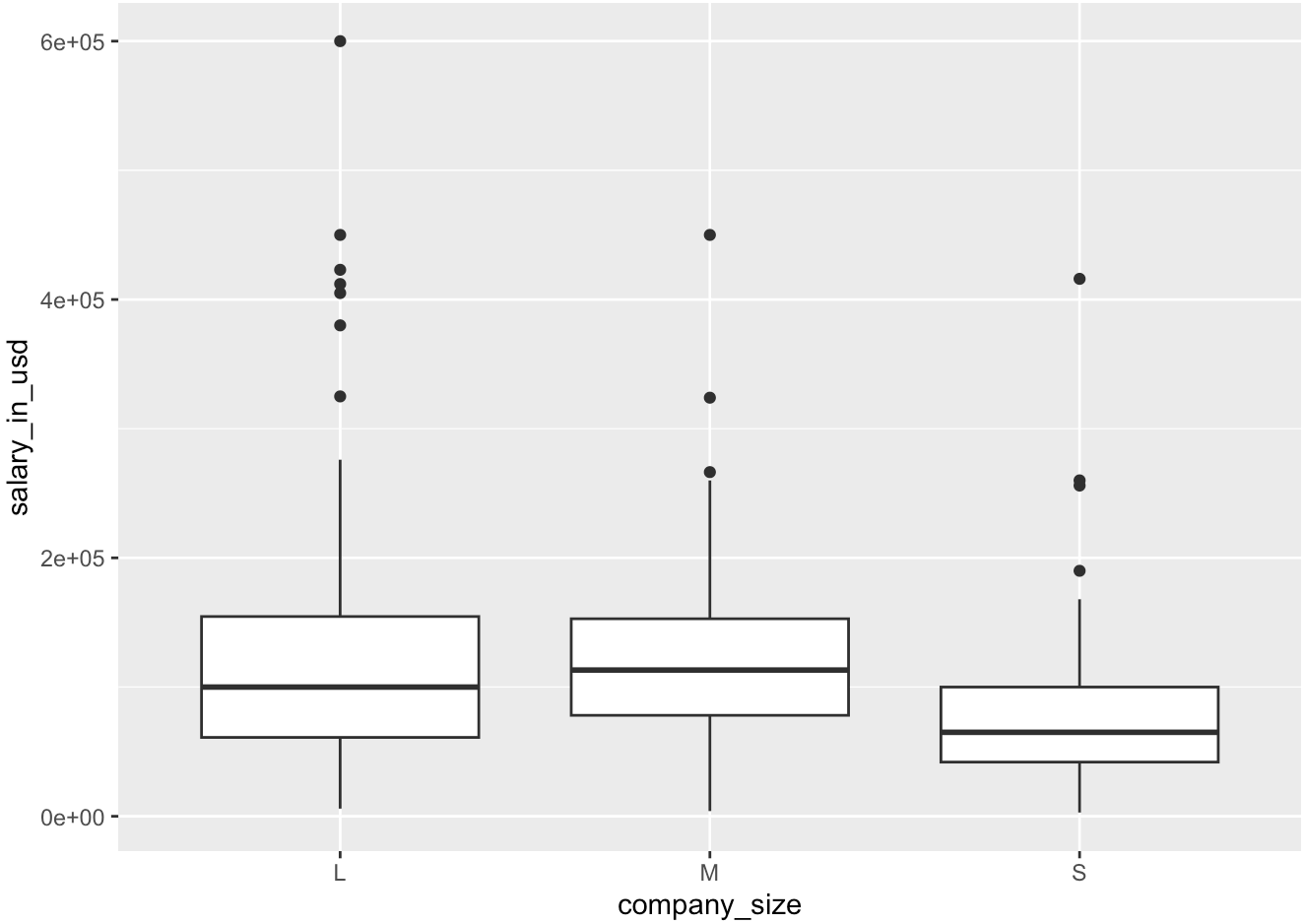
Plot for variable remote_ratio

```
ggplot(salaries, aes(x= remote_ratio, y = salary_in_usd))+
  geom_boxplot()
```



Plot for variable company_size

```
ggplot(salaries, aes(x= company_size , y = salary_in_usd))+
  geom_boxplot()
```



Analyzing Data:

1). What are the salaries of full time data scientist positions by experience over the years?

Filtering the data from Salaries of Full-Time Data Scientists in the US from 2020 to 2022.

```
require(dplyr)

  # Find only Full time employees
filtered_salaries<- salaries %>% filter(grepl("FT", employment_type))

  # Change the order of experience level and rename for nice graph display
filtered_salaries$experience_level <- factor(filtered_salaries$experience_level, levels
= c("EN", "MI", "SE","EX"))

 new_filtered_salaries<- filtered_salaries %>%
  mutate(experience_level = recode(experience_level,
                        "EN" = "Entry-level",
                        "MI" = "Mid-level",
                        "SE" = "Senior-level",
                        "EX" = "Executive-level"))

  # Identify and remove outliers for more narrowed down data
outliers <- boxplot.stats(new_filtered_salaries$salary_in_usd)$out

cleaned_salaries <- new_filtered_salaries[!new_filtered_salaries$salary_in_usd %in% outl
iers, ]

head(cleaned_salaries)
```

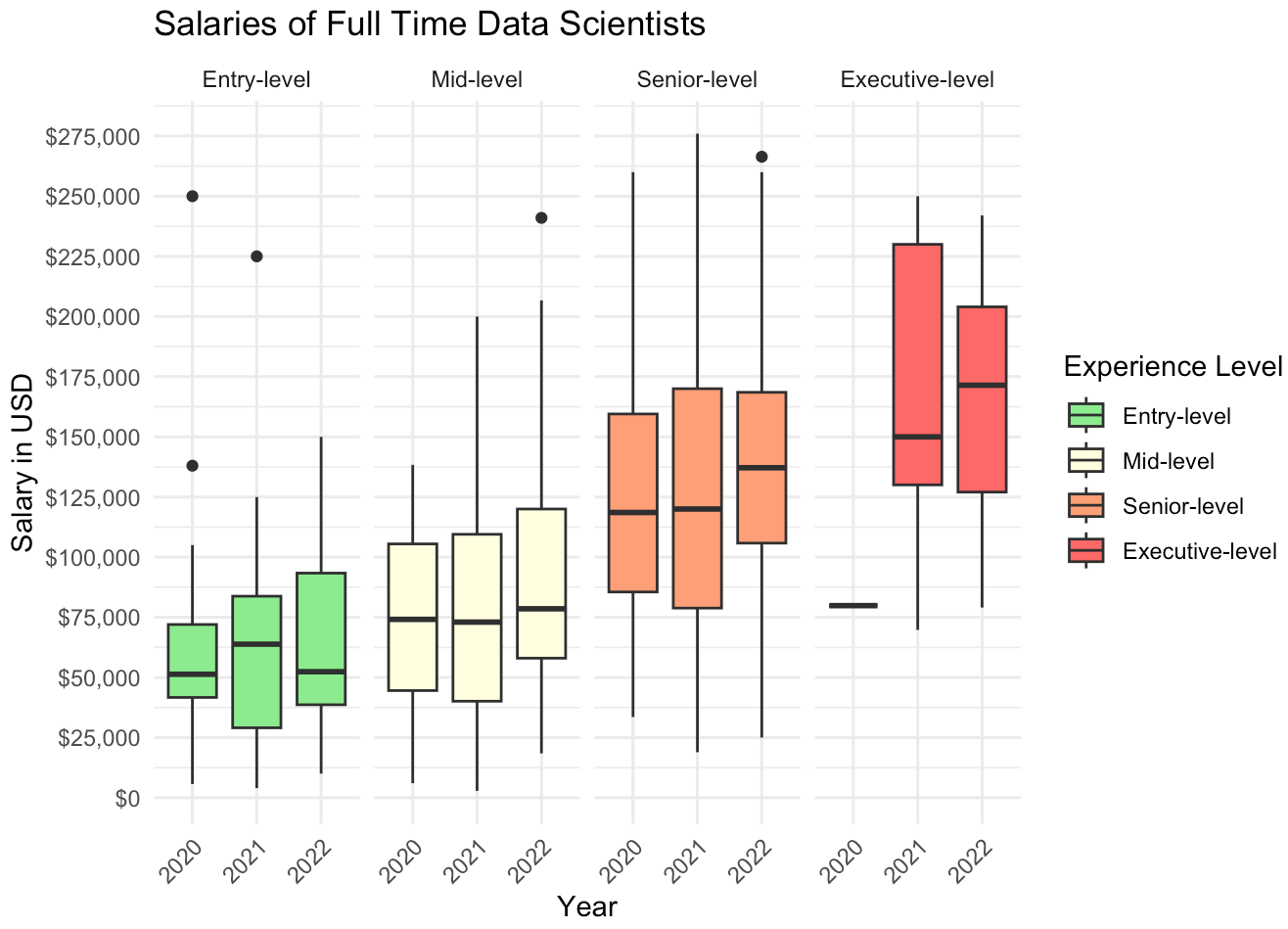
```
##   X work_year experience_level employment_type                 job_title
## 1 0      2020        Mid-level              FT             Data Scientist
## 2 1      2020     Senior-level              FT Machine Learning Scientist
## 3 2      2020     Senior-level              FT          Big Data Engineer
## 4 3      2020        Mid-level              FT       Product Data Analyst
## 5 4      2020     Senior-level              FT  Machine Learning Engineer
## 6 5      2020      Entry-level              FT               Data Analyst
##   salary salary_currency salary_in_usd employee_residence remote_ratio
## 1  70000             EUR         79833                 DE            0
## 2 260000             USD        260000                 JP            0
## 3  85000             GBP        109024                 GB           50
## 4  20000             USD         20000                 HN            0
## 5 150000             USD        150000                 US           50
## 6  72000             USD         72000                 US          100
##   company_location company_size
## 1               DE            L
## 2               JP            S
## 3               GB            M
## 4               HN            S
## 5               US            L
## 6               US            L
```

Graphing the data

```
require("scales")

custom_colors <- c("Entry-level" = "lightgreen", "Mid-level" = "lightyellow", "Senior-le
vel" = "lightsalmon","Executive-level" = "indianred1")

ggplot(cleaned_salaries, aes(x = work_year, y =salary_in_usd, fill = experience_level))+
  geom_boxplot()+
  facet_grid(.~experience_level)+
  labs(fill = "Experience Level")+  # Labeling the legend
  scale_fill_manual(values = custom_colors) + #adding color
  scale_y_continuous(labels = dollar, breaks = seq(0, 600000, by = 25000)) +
  labs(title = "Salaries of Full Time Data Scientists",
       x = "Year",
       y = "Salary in USD") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) #making x-axis lables angled
```



Analyze Results:

As you many already know, Data Scientist salaries vary wildly around the world and year to year. I'm going to simplify a salary range that will be perfect for paying your new Data Scientist employee. I'll make a note that Before graphing this data, I disregarded some big outliers so we can start narrowing down that range.

Let's look at the graph. On the x-axis we have Years from 2020 to 2022, on the y-axis we have Salaries in USD of all full time Data Scientists, and the data is grouped and colored by experience levels, from Entry to Executive level. Sure enough, we can see that as the years increase within each experience level, the majority of the salaries also increase. I did some further research and found that the employment of data scientists is projected to grow 36% from 2021 to 2031 (Rutgers University, 2024). We must take this into account so we can have an accurate range for today's salaries. You can also see that as you move from an Entry-level position to the Executive-level position, the salaries increase

Ham, C., Hann, R. N., Wang, W., & Yang, J. (2023). Going remote? the role of Labor Market Competition. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.4201819 (https://doi.org/10.2139/ssrn.4201819)

2). For the estimated 2025 salary percent increase, what are the Median and Interquartile salary ranges of full time data scientists by experience levels?

```
require("knitr")
require("kableExtra")

# Function to add a percentage to a number for the projected 36% increase from 2021 to 2
031, with 14.4% from 2021 to today 2025

add_percentage <- function(original_number, percentage) {
  increase <- original_number * (percentage / 100)
  new_number <- original_number + increase
  return(new_number)
}

#Creating a data frame of Median and Interquartile salary ranges

salaries_data <- cleaned_salaries%>%
  group_by(experience_level) %>%
  summarize(
    Median= median(salary_in_usd),
    Q1 = paste(quantile(salary_in_usd, 0.25)),
    Q3 = paste(quantile(salary_in_usd, 0.75)), .groups ='drop')

 #Renaming the columns

colnames(salaries_data) <- c("Experience Level", "Median", "Q1", "Q3")

 #Making the values numeric

salaries_data$Q1 <- as.numeric(salaries_data$Q1)
salaries_data$Q3<- as.numeric(salaries_data$Q3)

# Using function to add a percentage to a number for the projected 36% increase from 202
1 to 2031, with 14.4% from 2021 to today 2025

salaries_data$Median <- add_percentage(salaries_data$Median, 14.4)
salaries_data$Q1<- add_percentage(salaries_data$Q1, 14.4)
salaries_data$Q3<- add_percentage(salaries_data$Q3, 14.4)


  #Adding $ to the values

salaries_data$Median <- dollar(salaries_data$Median)
salaries_data$Q1<- dollar(salaries_data$Q1)
salaries_data$Q3<- dollar(salaries_data$Q3)

#Creating a presentable table

kable(salaries_data, caption = "Estimated 2025 Full time Data Scientist Salaries") %>%
  kable_styling("striped", full_width = FALSE, position = "center", font_size = 12)%>%
  row_spec(0, background = "grey84", color = "black", bold = TRUE)%>%
  row_spec(1:1, background = "lightgreen", color = "black")%>%
  row_spec(2:2, background = "lightyellow", color = "black")%>%
  row_spec(3:3, background = "lightsalmon", color = "black")%>%
```

```
  row_spec(4:4, background = "indianred1", color = "black")%>%
  column_spec(1:4, border_left = TRUE, border_right = TRUE)
```

Estimated 2025 Full time Data Scientist Salaries

Experience Level	Median	Q1	Q3
Entry-level	$67,613	$38,365	$98,214
Mid-level	$88,019	$55,748	$126,855
Senior-level	$154,440	$114,400	$194,480
Executive-level	$173,698	$138,584	$242,528

Analyze Results:

Let's look more deeply and see what the projected 2025 full time Data Scientist salaries would be by looking at a table. This table shows the estimated 2025 salaries. The table contains the Median and the Interquartile range of the middle 50% of our data. It goes from Q1, the lower part of that range, and Q3, the upper part of that range. This way we can see the majority of our data and narrow down the range even further.

Since you currently have a small company and want your data scientist growing and learning the ways of your company as it continues to expand rapidly, you should look at hiring a data scientist that is at the Entry to Mid Level position. This way they will advance to the Senior or Executive level positions by the time your company becomes large, being able to lead new employees in the future. Also, to be competitive and to get top talent, it is important to look more toward the higher end of the salaries range, looking at the range from the median, up to Q3.

3).For the estimated 2025 salary percent increase, what are the Entry and Mid level salaries of full time data scientists in the US vs not in the US among the different company sizes?

Filtering data so we can compare company location in US vs not in the US.

```
 #filtering employee location for in US or Not in US and experience level for Entry and
Mid levels

salaries_location <- cleaned_salaries %>%
  mutate(employee_residence = ifelse(grepl("US", company_location), company_location, "N
ot in US"))%>%
  filter(experience_level %in% c("Entry-level","Mid-level" ))


#Reordering and renaming company size from Small to Large for nice graph purposes

salaries_location$company_size <- factor(salaries_location$company_size, levels = c("S",
"M", "L"))

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


# Using function to add a percentage to a number for the projected 36% increase from 202
1 to 2031, with 14.4% from 2021 to today 2025

new_salaries_location$salary_in_usd <- add_percentage(new_salaries_location$salary_in_us
d, 14.4)


  # Identify and remove outliers for more narrowed down data

outliers <- boxplot.stats(new_salaries_location$salary_in_usd)$out

cleaned_locations <- new_salaries_location[!new_salaries_location$salary_in_usd %in% out
liers, ]

head(cleaned_locations)
```

```
##    X work_year experience_level employment_type            job_title   salary
## 1  0      2020        Mid-level              FT        Data Scientist    70000
## 2  3      2020        Mid-level              FT  Product Data Analyst    20000
## 3  5      2020      Entry-level              FT          Data Analyst    72000
## 4  7      2020        Mid-level              FT        Data Scientist 11000000
## 5  8      2020        Mid-level              FT Business Data Analyst   135000
## 6 10      2020      Entry-level              FT        Data Scientist    45000
##   salary_currency salary_in_usd employee_residence remote_ratio
## 1             EUR      91328.95          Not in US            0
## 2             USD      22880.00          Not in US            0
## 3             USD      82368.00                 US          100
## 4             HUF      40880.84          Not in US           50
## 5             USD     154440.00                 US          100
## 6             EUR      58711.22          Not in US            0
##   company_location company_size
## 1               DE        Large
## 2               HN        Small
## 3               US        Large
## 4               HU        Large
## 5               US        Large
## 6               FR        Small
```

Making a graph of the data to compare company location in US vs off shore. Y-axis is salary in USD and x-axis is company size, colored by employee location.

```
ggplot(cleaned_locations , aes( x = company_size, y= salary_in_usd, fill = employee_resi
dence)) +
  geom_boxplot()+
  scale_fill_manual(values = c("Not in US" = "lightblue", "US" = "sienna1"))+
  labs(fill = "Employee Location")+
  scale_y_continuous(labels = dollar, breaks = seq(0, 250000, by = 15000)) +
  labs(title = "Estimated 2025 Entry and Mid Level Salaries",
       x = "Company Size",
       y = "Salary in USD") +
  theme_minimal()
```

Analyze Results:

When graphing the Entry and Mid level positions, I also disregarded some big outliers so we can get a narrowed down range. This graph shows the estimated 2025 salaries.

Let's start by looking at this graph. You will notice that I combined the Entry and Mid level positions. On the x-axis, we can see the company size, where Small represents less than 50 employees, Medium represents 50 to 250 employees, and Large represents more than 250 employees. The salaries in USD are on the y-axis and the graph is colored by the location of the employee, either in the US or not.

We can see that employees located in the US pay their Entry to Mid level employees a higher salary than those that are not located in US. We can also see that as the company grows from a small company to a large company, the salaries tend to increase from small to medium, and level out, slightly decreasing from medium to large. That makes sense when we think about it. Small companies aren't making as big as a profit so their employees salaries are lower, where medium sized companies have some more employees but are making a larger profit so they can pay their employees a slightly higher salary, and large companies have so many employees and their profits might not be enough to raise employee salaries, so they are paying their employees about the same or less than the medium sized companies.

Since your company is small at the moment, let's narrow down our range even more by looking closer at table of this data for smaller companies.

4). For the estimated 2025 salary percent increase, what are the Median and Interquartile salary ranges of Entry & Mid level full time data scientists salaries among small companies by location?

```
#Creating a data frame of Median and Interquartile salary ranges while only looking at t
he small companies

small_company <- cleaned_locations %>%
  filter(company_size == "Small")%>%
  group_by(employee_residence, experience_level) %>%
  summarize(
    Median= median(salary_in_usd),
    Q1 = paste(quantile(salary_in_usd, 0.25)),
    Q3 = paste(quantile(salary_in_usd, 0.75)), .groups ='drop')

 #Renaming the column names

colnames(small_company) <- c("Employee Location", "Experience Level", "Median", "Q1", "Q
3")

 #Making the values numeric

small_company$Q1 <- as.numeric(small_company$Q1)
small_company$Q3<- as.numeric(small_company$Q3)


 #Adding $ to the values

small_company$Median <- dollar(small_company$Median)
small_company$Q1<- dollar(small_company$Q1)
small_company$Q3<- dollar(small_company$Q3)


  #Creating a presentable table

kable(small_company , caption = "Estimated 2025 Small Company Data Scientist Salaries")
%>%
  kable_styling("striped", full_width = FALSE, position = "center", font_size = 12) %>%
  row_spec(0, background = "grey84", color = "black", bold = TRUE)%>%
  row_spec(1:2, background = "lightblue", color = "black")%>%
  row_spec(3:4, background = "sienna1", color = "black")%>%
  column_spec(1:5, border_left = TRUE, border_right = TRUE)
```

Estimated 2025 Small Company Data Scientist Salaries

Employee Location	Experience Level	Median	Q1	Q3
Not in US	Entry-level	$54,434	$21,209.47	$83,862
Not in US	Mid-level	$51,182	$22,308.00	$73,546
US	Entry-level	$102,960	$83,512.00	$117,260
US	Mid-level	$66,352	$54,912.00	$109,533

Analyze Results:

Now we are looking at estimated salaries for 2025 of Full time Data Scientists that are at Entry to Mid level positions, working at small companies, with the employees located in and out of the US. This table shows the Median and Interquartile range, from Q1 to Q3, similar to the previous table.

Let's look at the higher end of the salaries, from the Median to Q3 again to be competitive and to get top talent.

With the projected 36% increase from 2021 to 2031, if you were to hire a Entry to Mid level employee not located in the US today, you are looking to pay them anywhere from $51,000 to $84,000. If you were to hire a Entry to Mid level employee working in the US today, you are looking to pay them anywhere from $66,000 to $117,000.

Conclusion:

So, I would recommend paying a data scientist anywhere from **$66,000 to $117,000**. This would get you a full-time data scientist that is at an Entry or Mid level position that can grow with your company and eventually become someone who can drive data science within the entire organization and could potentially lead a team in the future. This salary range is for a data scientist that is used to working for a small company and in the US. If you decide you don't mind your employee working outside of the US, you can pay them anywhere from **$51,000 to $84,000**. All these ranges account for the higher salary ranges due to the market being highly competitive. This higher salary range will also appeal to top talented data scientists.