R Project

Stephanie Halsing

2023-11-26

Step 1: Load R packages and read csv file.

```
# Load packages
library(readr)
library(dplyr)

##
## Attaching package: 'dplyr'

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

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

library(tidyr)
library(ggplot2)

# Read csv file to df
salaries <-
read_csv('https://raw.githubusercontent.com/lowhornj/DSE5002/main/R%20Project
/data.csv')

## New names:
## • `` -> `...1`

## Rows: 607 Columns: 12
## ── Column specification
────────────────────────────────────────────────────────
## Delimiter: ","
## chr (7): experience_level, employment_type, job_title, salary_currency,
empl...
## dbl (5): ...1, work_year, salary, salary_in_usd, remote_ratio
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this
message.

head(salaries)
```

```
## # A tibble: 6 × 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>
```

Step 2: Data wrangling - tidy, clean and organize data.

```r
# Replace name for first column
colnames(salaries)[1] <- c("Number")

# Create df for FT employees residing in the US
salaries_us <- salaries %>%
  filter(employment_type == "FT", employee_residence == "US") %>%
  select(work_year:job_title, salary_in_usd:company_size)

# Create df for FT employees residing in places other than the US
salaries_intl <- salaries %>%
  filter(employment_type == "FT", employee_residence != "US") %>%
  select(work_year:job_title, salary_in_usd:company_size)

# Replace country names with 'Intl' for the employee residences
salaries_intl$employee_residence <- c("Intl")

# Create df for job title containing 'Lead'
salaries_lead <- salaries %>%
  filter(grepl('Lead', job_title)) %>%
  select(work_year:job_title, salary_in_usd:company_size)
```

Step 3: Analysis (3 parts)

```
    A. Perform calculations for the each dataframe: original file, US
employees, International (Intl) employees and job titles containing 'Lead'
    B. Analyze salary differences between onsite and offshore employees (US
```

```
vs Intl employees)
    C. Analyze salaries between different company sizes (US employees only)

########Part A########

# Original file
summary(salaries$salary_in_usd)  # Calculate summary statistics

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2859   62726  101570  112298  150000  600000

print(IQR(salaries$salary_in_usd))  # Calculate IQR

## [1] 87274

# Calculate average salary for each experience level
print(aggregate(salary_in_usd ~ experience_level, data = salaries, mean))

##   experience_level salary_in_usd
## 1               EN      61643.32
## 2               EX     199392.04
## 3               MI      87996.06
## 4               SE     138617.29

# Filter by FT then find range of salaries for each experience level
salaries %>%
  filter(employment_type == 'FT') %>%
  group_by(experience_level) %>%
  summarize(min(salary_in_usd), max(salary_in_usd))

## # A tibble: 4 × 3
##   experience_level `min(salary_in_usd)` `max(salary_in_usd)`
##   <chr>                           <dbl>                <dbl>
## 1 EN                               4000               250000
## 2 EX                              69741               600000
## 3 MI                               2859               450000
## 4 SE                              18907               412000

# US FT employees
summary(salaries_us$salary_in_usd)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   25000  106195  138475  148297  174250  600000

print(IQR(salaries_us$salary_in_usd))

## [1] 68055

# Calculate average salary for each experience level
print(aggregate(salary_in_usd ~ experience_level, data = salaries_us, mean))

##   experience_level salary_in_usd
## 1               EN      98660.71
```

```
## 2                EX     238133.93
## 3                MI     133849.61
## 4                SE     154154.77
```

```r
# Find range of salaries for each experience level
salaries_us %>%
  group_by(experience_level) %>%
  summarize(min(salary_in_usd), max(salary_in_usd))
```

```
## # A tibble: 4 × 3
##   experience_level `min(salary_in_usd)` `max(salary_in_usd)`
##   <chr>                           <dbl>                <dbl>
## 1 EN                              50000               250000
## 2 EX                             110000               600000
## 3 MI                              37236               450000
## 4 SE                              25000               412000
```

```r
# Intl FT employees
summary(salaries_intl$salary_in_usd)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2859   40408   63760   69530   88654  260000
```

```r
print(IQR(salaries_intl$salary_in_usd))
```

```
## [1] 48246
```

```r
# Calculate average salary for each experience level
print(aggregate(salary_in_usd ~ experience_level, data = salaries_intl,
mean))
```

```
##   experience_level salary_in_usd
## 1               EN      45679.20
## 2               EX     130392.55
## 3               MI      61834.48
## 4               SE      92284.41
```

```r
# Find range of salaries for each experience level
salaries_intl %>%
  group_by(experience_level) %>%
  summarize(min(salary_in_usd), max(salary_in_usd))
```

```
## # A tibble: 4 × 3
##   experience_level `min(salary_in_usd)` `max(salary_in_usd)`
##   <chr>                           <dbl>                <dbl>
## 1 EN                               4000               150000
## 2 EX                              69741               230000
## 3 MI                               2859               200000
## 4 SE                              18907               260000
```

```r
# Job title containing 'Lead'
summary(salaries_lead$salary_in_usd)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   19609   87233  116594  139533  167500  405000
```

```r
print(IQR(salaries_lead$salary_in_usd))
```

```
## [1] 80267
```

```r
# Calculate average salary for each experience level
print(aggregate(salary_in_usd ~ experience_level, data = salaries_lead,
mean))
```

```
##   experience_level salary_in_usd
## 1               EX     118187.00
## 2               MI      69402.25
## 3               SE     173073.56
```

```r
# Find range of salaries for each experience level
salaries_lead %>%
  group_by(experience_level) %>%
  summarize(min(salary_in_usd), max(salary_in_usd))
```

```
## # A tibble: 3 × 3
##   experience_level `min(salary_in_usd)` `max(salary_in_usd)`
##   <chr>                           <dbl>                <dbl>
## 1 EX                             118187               118187
## 2 MI                              19609               115000
## 3 SE                              40570               405000
```

```r
########Part B########

# Calculate average salary for each remote ratio (US employees)
salaries_us %>%
  group_by(remote_ratio) %>%
  summarize(average = mean(salary_in_usd))
```

```
## # A tibble: 3 × 2
##   remote_ratio average
##          <dbl>   <dbl>
## 1            0 141253.
## 2           50 139837.
## 3          100 150821.
```

```r
# Calculate average salary for each remote ratio (Intl employees)
salaries_intl %>%
  group_by(remote_ratio) %>%
  summarize(average = mean(salary_in_usd))
```

```
## # A tibble: 3 × 2
##   remote_ratio average
##          <dbl>   <dbl>
## 1            0  68187.
```

```
## 2           50  71883.
## 3          100  68758.
```

```r
# Calculate average salary for each experience level for US employees in US-based companies
salaries %>%
  filter(employee_residence == "US", company_location == "US") %>%
  group_by(experience_level) %>%
  summarize(aveage = mean(salary_in_usd))
```

```
## # A tibble: 4 × 2
##   experience_level  aveage
##   <chr>              <dbl>
## 1 EN                98707.
## 2 EX               249992.
## 3 MI               135618.
## 4 SE               153591.
```

```r
# Calculate average salary for each experience level for Intl employees in US-based companies
salaries %>%
  filter(employee_residence != "US", company_location == "US") %>%
  group_by(experience_level) %>%
  summarize(aveage = mean(salary_in_usd))
```

```
## # A tibble: 4 × 2
##   experience_level  aveage
##   <chr>              <dbl>
## 1 EN                12000
## 2 EX               150000
## 3 MI                67511.
## 4 SE               103614.
```

```r
# Job title containing 'Lead'
# Calculate average salary for each remote ratio
salaries_lead %>%
  group_by(remote_ratio) %>%
  summarize(average = mean(salary_in_usd))
```

```
## # A tibble: 3 × 2
##   remote_ratio average
##          <dbl>   <dbl>
## 1            0 159644
## 2           50 108523.
## 3          100 143620.
```

```r
# Calculate average salary for each company location
salaries_lead %>%
  group_by(company_location) %>%
  summarize(average = mean(salary_in_usd))
```

```
## # A tibble: 7 × 2
##   company_location average
##   <chr>              <dbl>
## 1 AE               115000
## 2 CA               118187
## 3 DE                87932
## 4 GB               103160
## 5 IN                30090.
## 6 NZ               125000
## 7 US               192000
```

```
########Part C########

# Calculate average salary for each company size
salaries %>%
  group_by(company_size) %>%
  summarize(average = mean(salary_in_usd))
```

```
## # A tibble: 3 × 2
##   company_size average
##   <chr>          <dbl>
## 1 L            119243.
## 2 M            116905.
## 3 S             77633.
```

```
# Calculate average salary for each company size (US-based employees)
salaries_us %>%
  group_by(company_size) %>%
  summarize(average = mean(salary_in_usd))
```

```
## # A tibble: 3 × 2
##   company_size average
##   <chr>          <dbl>
## 1 L            167752.
## 2 M            143301.
## 3 S            105425
```

```
# Calculate average salary for each company size (Intl-based employees)
salaries_intl %>%
  group_by(company_size) %>%
  summarize(average = mean(salary_in_usd))
```

```
## # A tibble: 3 × 2
##   company_size average
##   <chr>          <dbl>
## 1 L             70060.
## 2 M             70753.
## 3 S             66329.
```

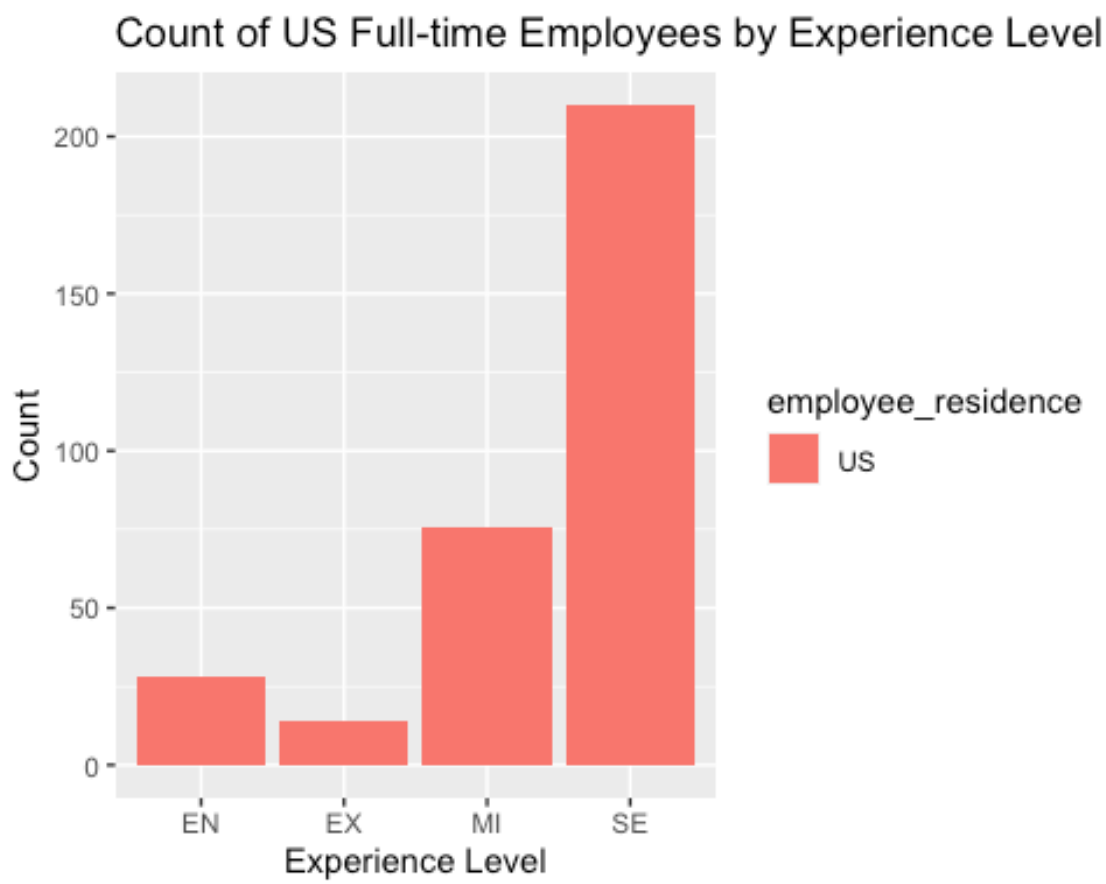
```
# Calculate average salary for each company size (job title containing
'Lead')
```

```r
salaries_lead %>%
  group_by(company_size) %>%
  summarize(average = mean(salary_in_usd))
```

```
## # A tibble: 3 × 2
##   company_size average
##   <chr>          <dbl>
## 1 L            159026.
## 2 M             71966
## 3 S            139269.
```

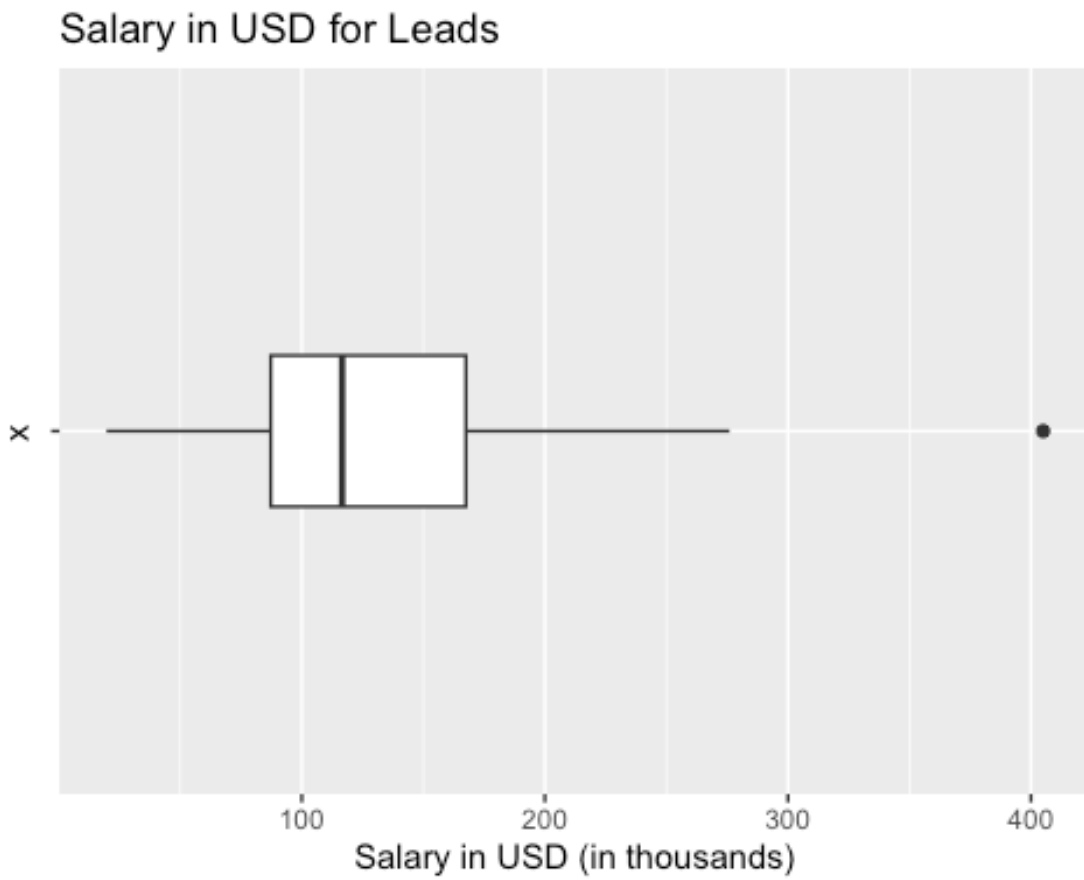
Step 4: Plots (3 parts)

```r
    A. Plot employee and salary information for US employees and job titles
containing 'Lead'
    B. Plot salary differences between US and Intl employees and remote
ratios for job titles containing 'Lead'
    C. Plot salary differences between different company sizes for US
employees

########Part A########

# Plot count of US FT employees for each experience level
ggplot(salaries_us, aes(x=experience_level, fill=employee_residence)) +
  geom_bar() + labs(x='Experience Level', y='Count',
                    title='Count of US Full-time Employees by Experience
Level')
```



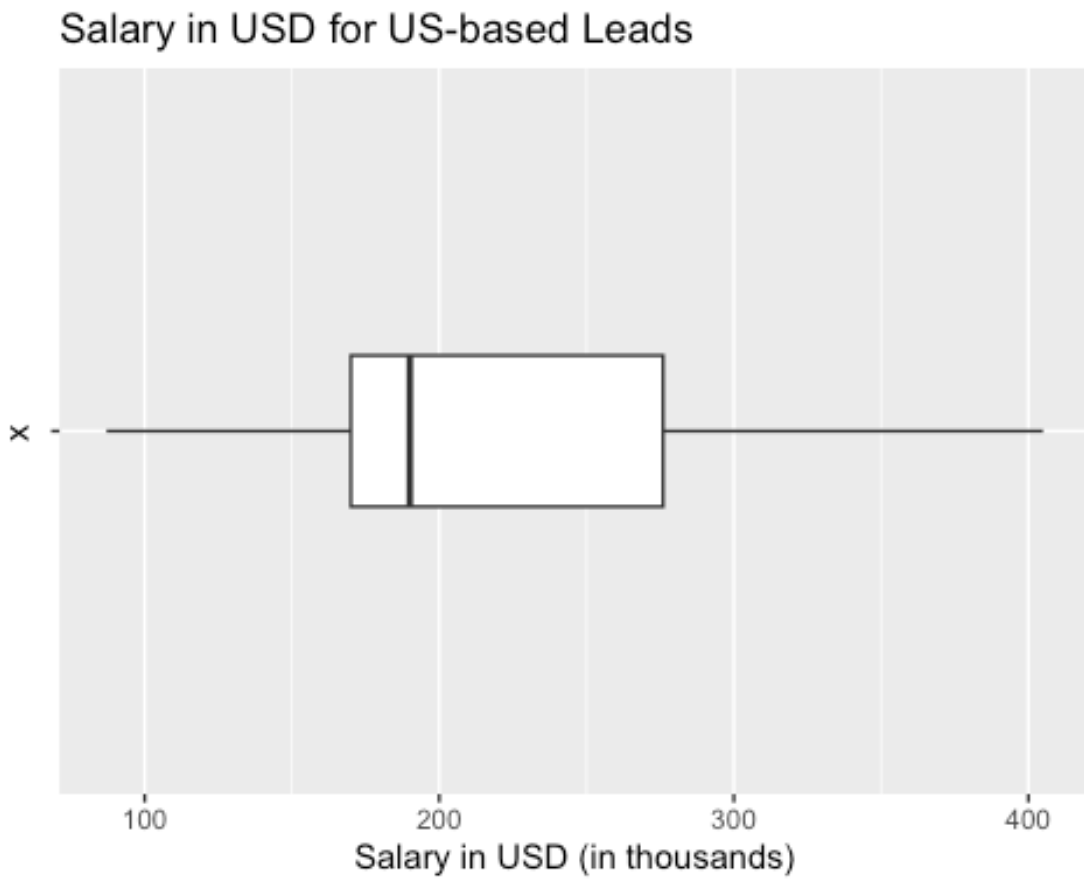

```
# Large portion of US employees are 'SE' experience level

# Plot boxplot for US FT employees
salaries_us %>%
  mutate(y_thousands = salary_in_usd/1e3) %>%
  ggplot(aes(x= '', y=y_thousands)) +
  geom_boxplot(width = 0.25) + coord_flip() +
  labs(x='x', y='Salary in USD (in thousands)', title='Salary in USD for US
Full-time employees')
```


```
# Plot boxplot for job title containing 'Lead'
salaries_lead %>%
  mutate(y_thousands = salary_in_usd/1e3) %>%
  ggplot(aes(x= '', y=y_thousands)) +
  geom_boxplot(width = 0.25) + coord_flip() +
  labs(x='x', y='Salary in USD (in thousands)', title='Salary in USD for
Leads')
```



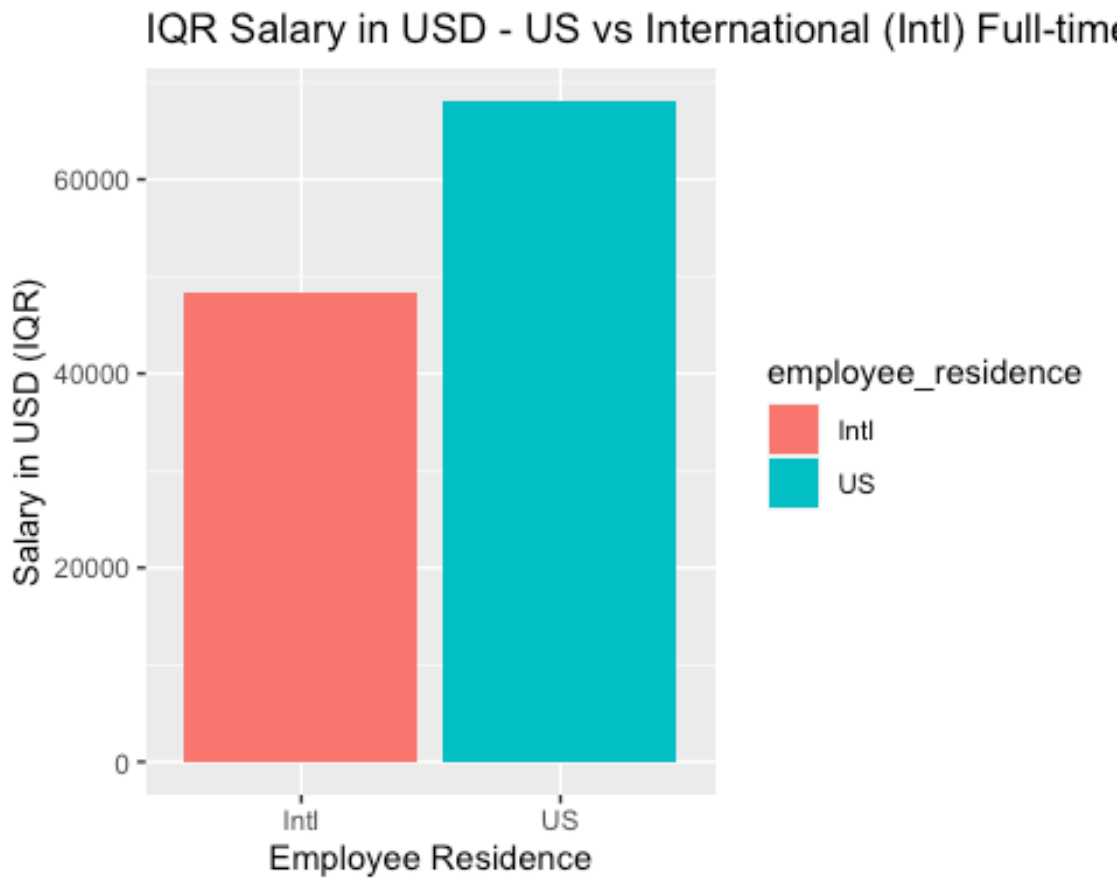

```
# Plot average salary per employee residence for job title containing 'Lead'
ggplot(salaries_lead, aes(x=employee_residence, fill=employee_residence,
y=salary_in_usd)) +
  geom_bar(stat = 'summary', fun = 'mean') +
  labs(x='Employee Residence', y='Salary in USD (Average)',
       title='Average Salary in USD per Employee Residence for Leads')
```


```
# Plot boxplot for job title containing 'Lead' for employees residing in the US
salaries_lead %>%
  filter(employee_residence == 'US') %>%
  mutate(y_thousands = salary_in_usd/1e3) %>%
  ggplot(aes(x= '', y=y_thousands)) +
  geom_boxplot(width = 0.25) + coord_flip() +
  labs(x='x', y='Salary in USD (in thousands)', title='Salary in USD for US-based Leads')
```

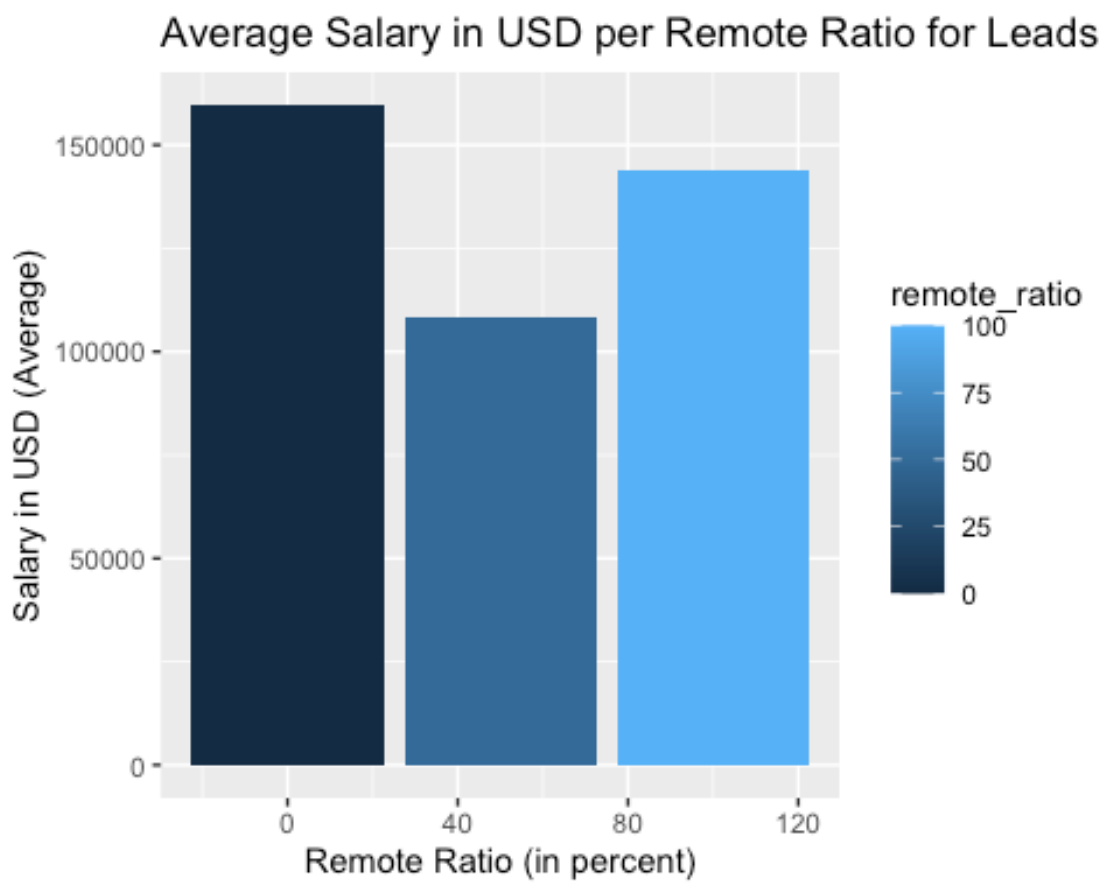



```
# Print summary statistics and IQR for the previous boxplot
salaries_lead2 <- salaries_lead %>%
  filter(employee_residence == 'US') %>%
  select(work_year:job_title, salary_in_usd:company_size)

summary(salaries_lead2$salary_in_usd)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   87000  170000  190000  225600  276000  405000

print(IQR(salaries_lead2$salary_in_usd))

## [1] 106000

########Part B########

# Combine US and Intl dfs
combined_plot <- rbind(salaries_us, salaries_intl)

# Plot IQR salary in USD for US and Intl FT employees
ggplot(combined_plot,
       aes(x=employee_residence, fill=employee_residence, y=salary_in_usd)) +
        geom_bar(stat = 'summary', fun = 'IQR') +
```

```
        labs(x='Employee Residence', y='Salary in USD (IQR)', title='IQR
Salary in USD - US vs International (Intl) Full-time Employees')
```




```
# Plot average salary per remote ratio for job title containing 'Lead'
ggplot(salaries_lead, aes(x=remote_ratio, fill=remote_ratio,
y=salary_in_usd)) +
  geom_bar(stat = 'summary', fun = 'mean') +
  labs(x='Remote Ratio (in percent)', y='Salary in USD (Average)',
        title='Average Salary in USD per Remote Ratio for Leads')
```




```
########Part C########

# Plot average salary per company size for US FT employees
ggplot(salaries_us, aes(x=company_size, fill=company_size, y=salary_in_usd))
+
  geom_bar(stat = 'summary', fun = 'mean') +
  labs(x='Company Size', y='Salary in USD (Average)',
       title='Average Salary in USD per Company Size for US Full-time
Employees')
```


```
# Plot average salary per company size for job title containing 'Lead'
ggplot(salaries_lead, aes(x=company_size, fill=company_size,
y=salary_in_usd)) +
  geom_bar(stat = 'summary', fun = 'mean') +
  labs(x='Company Size', y='Salary in USD (Average)',
       title='Average Salary in USD per Company Size for Leads')
```


Average Salary in USD per Company Size for Leads
Salary in USD (Average)
150000
100000
50000
0
L
M
S
Company Size
company_size
L
M
S