Salary Analysis

04/10/25

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(lubridate)
library(flextable)
```

```
##
## Attaching package: 'flextable'
##
## The following object is masked from 'package:purrr':
##
##     compose
```

```
library(ggplot2)
```

```
df <- read_csv("/Users/joshsoiro/Desktop/r\ project\ data-1.csv",
               col_types = cols(
                 work_year           = col_integer(),
                 experience_level    = col_character(),
                 employment_type     = col_character(),
                 job_title           = col_character(),
                 salary              = col_double(),
                 salary_currency     = col_character(),
                 salary_in_usd         = col_double(),
                 employee_residence  = col_character(),
                 remote_ratio        = col_double(),
                 company_location    = col_character(),
                 company_size        = col_character()
               ))
```

```
## New names:
## * `` -> `...1`
```

```
df <- df %>%
  mutate(
    experience_level = factor(experience_level, levels = c("EN","MI","SE","EX"), ordered = TRUE),
    employment_type  = factor(employment_type, levels = c("PT","FT","CT","FL")),
```

```
    remote_ratio     = factor(remote_ratio, levels = c(0,50,100), labels = c("On-site","Hybrid","Remote
    company_size     = factor(company_size, levels = c("S","M","L"), labels = c("Small","Medium","Large
    employee_residence = factor(employee_residence),
    company_location   = factor(company_location),
    us_flag = if_else(company_location == "US", "US", "Non-US")
  )
print(df)
```

```
## # A tibble: 607 x 13
##     ...1 work_year experience_level employment_type job_title             salary
##    <dbl>     <int> <ord>            <fct>           <chr>                  <dbl>
##  1     0      2020 MI               FT              Data Scientist        7   e4
##  2     1      2020 SE               FT              Machine Learning Sci~ 2.6 e5
##  3     2      2020 SE               FT              Big Data Engineer     8.5 e4
##  4     3      2020 MI               FT              Product Data Analyst  2   e4
##  5     4      2020 SE               FT              Machine Learning Eng~ 1.5 e5
##  6     5      2020 EN               FT              Data Analyst          7.20e4
##  7     6      2020 SE               FT              Lead Data Scientist   1.9 e5
##  8     7      2020 MI               FT              Data Scientist        1.10e7
##  9     8      2020 MI               FT              Business Data Analyst 1.35e5
## 10     9      2020 SE               FT              Lead Data Engineer    1.25e5
## # i 597 more rows
## # i 7 more variables: salary_currency <chr>, salary_in_usd <dbl>,
## #   employee_residence <fct>, remote_ratio <fct>, company_location <fct>,
## #   company_size <fct>, us_flag <chr>
```

```
overview <- df %>%
  summarise(
    n            = n(),
    mean_USD     = mean(salary_in_usd, na.rm=TRUE),
    median_USD   = median(salary_in_usd, na.rm=TRUE),
    sd_USD       = sd(salary_in_usd, na.rm=TRUE),
    p10_USD      = quantile(salary_in_usd, .10, na.rm=TRUE),
    p25_USD      = quantile(salary_in_usd, .25, na.rm=TRUE),
    p75_USD      = quantile(salary_in_usd, .75, na.rm=TRUE),
    p90_USD      = quantile(salary_in_usd, .90, na.rm=TRUE)
  )
print(overview)
```

```
## # A tibble: 1 x 8
##       n mean_USD median_USD sd_USD p10_USD p25_USD p75_USD p90_USD
##   <int>    <dbl>      <dbl>  <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
## 1   607  112298.     101570 70957.  33689.   62726  150000  200000
```

```
by_experience <- df %>%
  group_by(experience_level) %>%
  summarise(
    count       = n(),
    mean_USD    = mean(salary_in_usd, na.rm=TRUE),
    median_USD  = median(salary_in_usd, na.rm=TRUE)
  )
print(by_experience)
```

```
## # A tibble: 4 x 4
##   experience_level count mean_USD median_USD
```

```
##   <ord>            <int>    <dbl>      <dbl>
## 1 EN                  88   61643.     56500
## 2 MI                 213   87996.     76940
## 3 SE                 280  138617.    135500
## 4 EX                  26  199392.    171438.
```

```
by_us <- df %>%
  group_by(us_flag) %>%
  summarise(
    count       = n(),
    mean_USD    = mean(salary_in_usd, na.rm=TRUE),
    median_USD  = median(salary_in_usd, na.rm=TRUE)
  )
print(by_us)
```

```
## # A tibble: 2 x 4
##   us_flag count mean_USD median_USD
##   <chr>   <int>    <dbl>      <dbl>
## 1 Non-US    252   67560.     62688.
## 2 US        355  144055.    135000
```

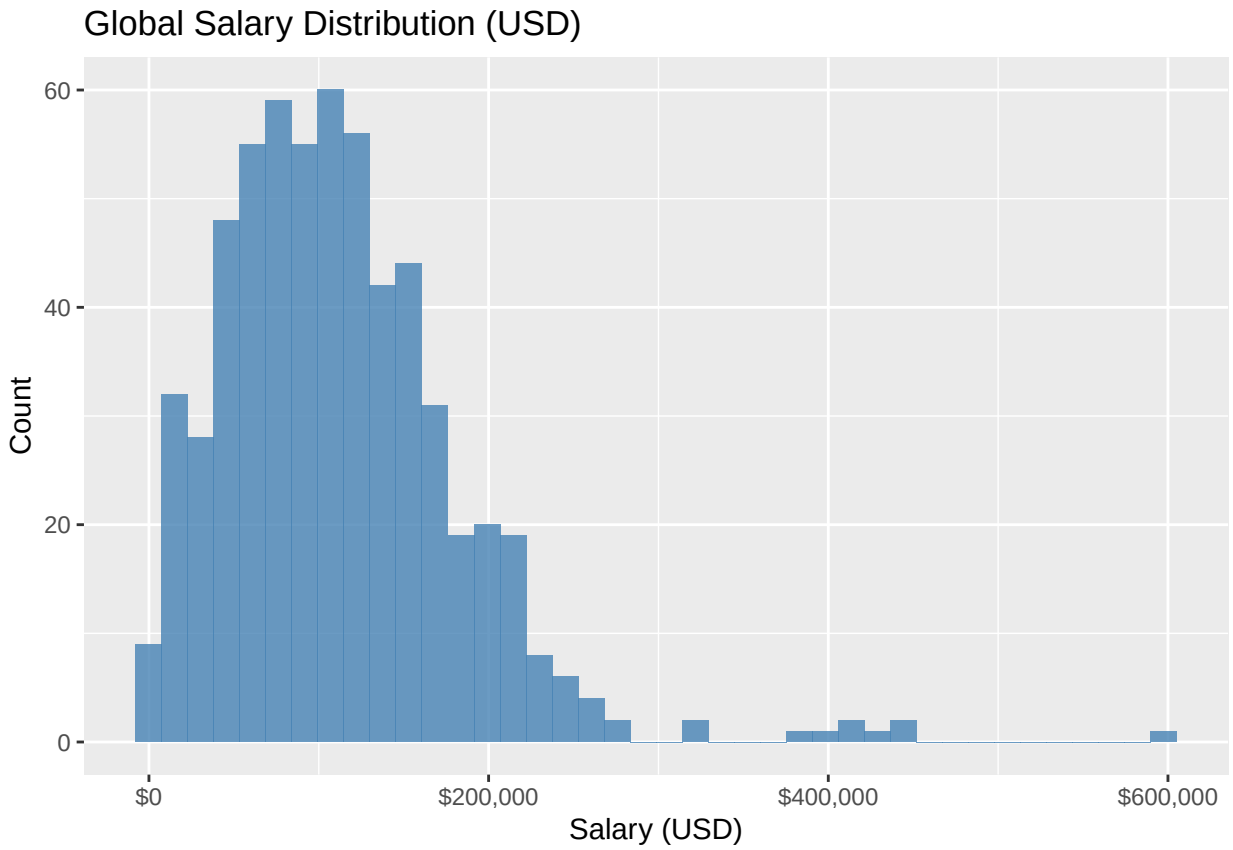
```
by_size <- df %>%
  group_by(company_size) %>%
  summarise(
    count      = n(),
    median_USD = median(salary_in_usd, na.rm=TRUE)
  )
print(by_size)
```

```
## # A tibble: 3 x 3
##   company_size count median_USD
##   <fct>        <int>      <dbl>
## 1 Small           83      65000
## 2 Medium         326     113188
## 3 Large          198     100000
```

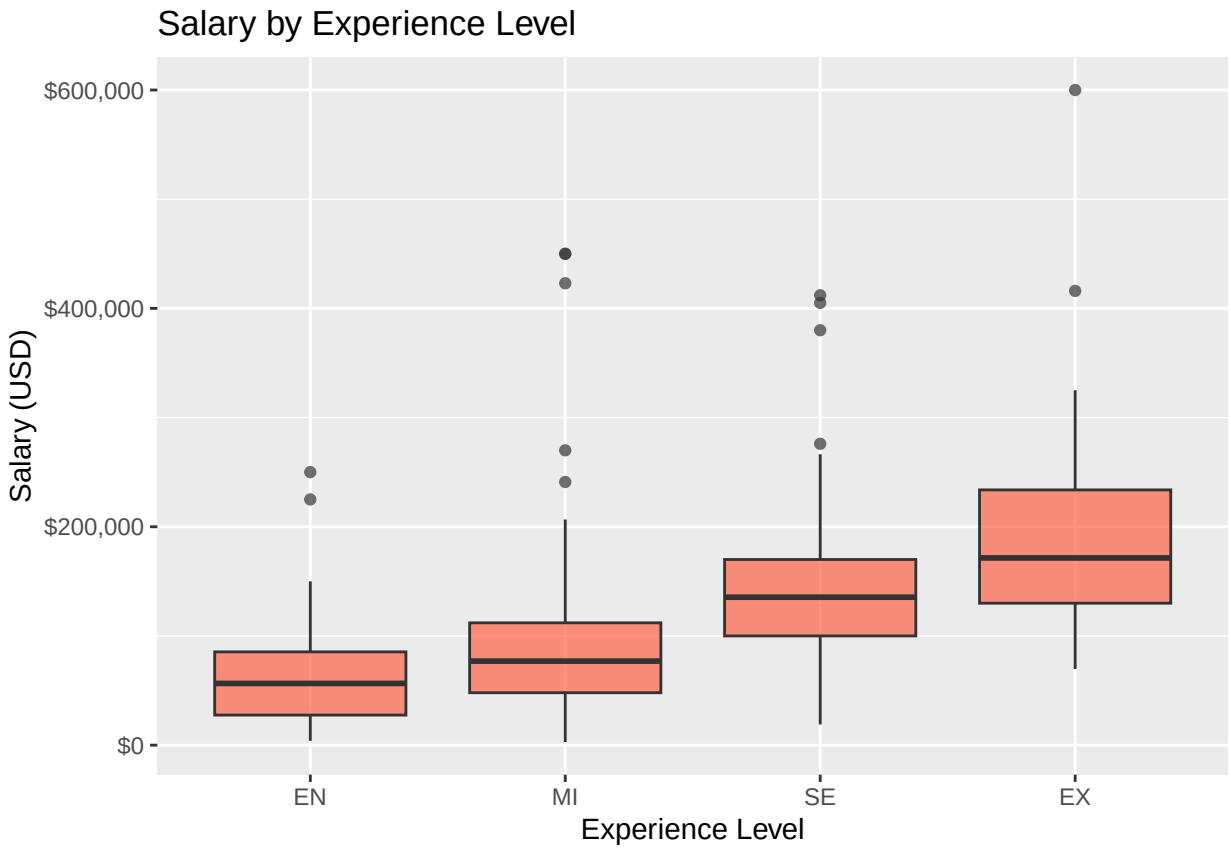
```
by_remote <- df %>%
  group_by(remote_ratio) %>%
  summarise(
    count      = n(),
    median_USD = median(salary_in_usd, na.rm=TRUE)
  )
print(by_remote)
```

```
## # A tibble: 3 x 3
##   remote_ratio count median_USD
##   <fct>        <int>      <dbl>
## 1 On-site        127      99000
## 2 Hybrid          99      69999
## 3 Remote         381     115000
```

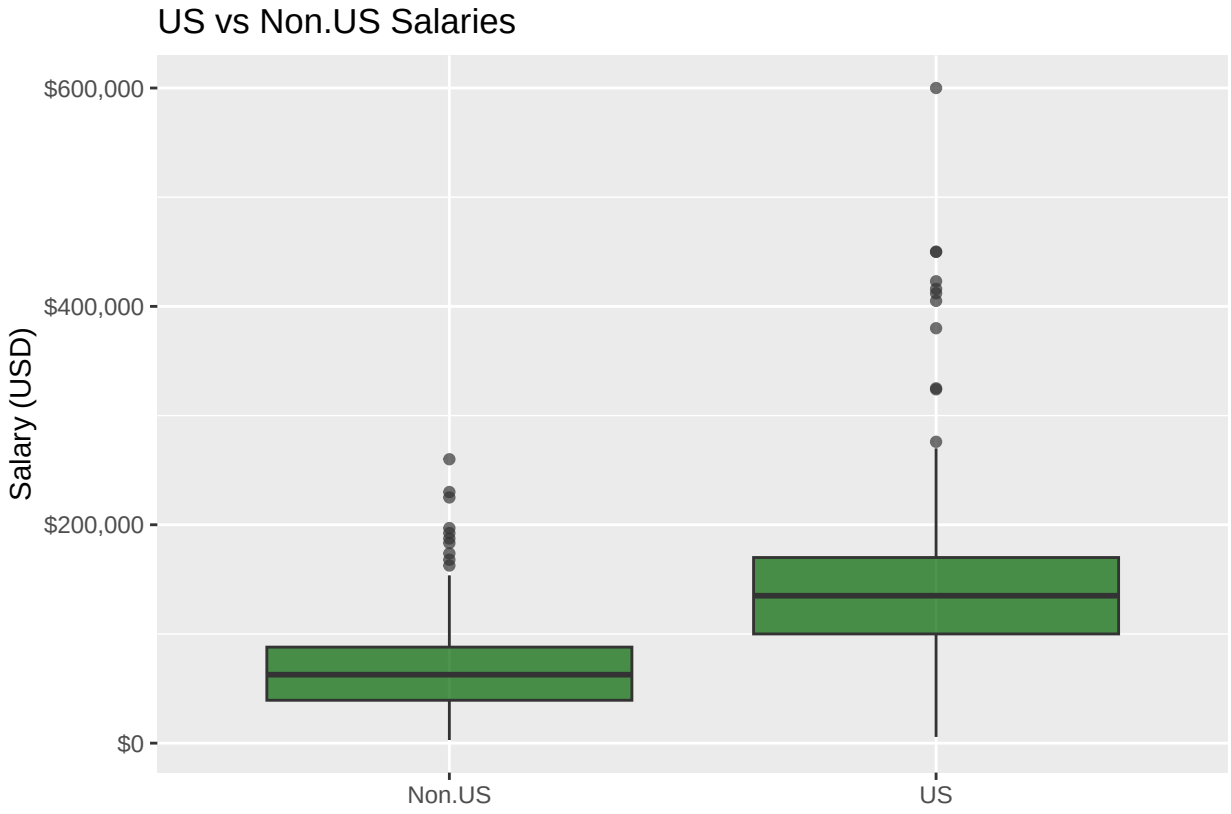
```
p_dist <- ggplot(df, aes(x = salary_in_usd)) +
  geom_histogram(bins = 40, alpha = .8, fill = "steelblue") +
  scale_x_continuous(labels = scales::dollar) +
  labs(title = "Global Salary Distribution (USD)", x = "Salary (USD)", y = "Count")
print(p_dist)
```




```
p_exp <- ggplot(df, aes(x = experience_level, y = salary_in_usd)) +
  geom_boxplot(fill = "tomato", alpha = .7) +
  scale_y_continuous(labels = scales::dollar) +
  labs(title = "Salary by Experience Level", x = "Experience Level", y = "Salary (USD)")
print(p_exp)
```




```
p_us <- ggplot(df, aes(x = us_flag, y = salary_in_usd)) +
  geom_boxplot(fill = "darkgreen", alpha = .7) +
  scale_y_continuous(labels = scales::dollar) +
  labs(title = "US vs Non-US Salaries", x = "", y = "Salary (USD)")
print(p_us)
```


US vs Non.US Salaries
$600,000
$400,000
$200,000
$0
Salary (USD)
Non.US
US