# Brian Wrenn Project 1

Brian Wrenn

2025-11-30

```
library(tidyverse)
```

```
## Warning: package 'tidyr' was built under R version 4.4.3
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✓ dplyr     1.1.4     ✓ readr     2.1.5
## ✓ forcats   1.0.0     ✓ stringr   1.5.1
## ✓ ggplot2   3.5.1     ✓ tibble    3.2.1
## ✓ lubridate 1.9.3     ✓ tidyr     1.3.1
## ✓ purrr     1.0.2
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✕ dplyr::filter() masks stats::filter()
## ✕ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becom
e errors
```

```
# CSV
ds <- read_csv(file.choose())
```

```
## 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`
```

```
glimpse(ds)
```

```
## Rows: 607
## Columns: 12
## $ ...1               <dbl> 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 1…
## $ work_year          <dbl> 2020, 2020, 2020, 2020, 2020, 2020, 2020, 2020, 202…
## $ experience_level   <chr> "MI", "SE", "SE", "MI", "SE", "EN", "SE", "MI", "MI…
## $ employment_type    <chr> "FT", "FT", "FT", "FT", "FT", "FT", "FT", "FT", "FT…
## $ job_title          <chr> "Data Scientist", "Machine Learning Scientist", "Bi…
## $ salary             <dbl> 70000, 260000, 85000, 20000, 150000, 72000, 190000,…
## $ salary_currency    <chr> "EUR", "USD", "GBP", "USD", "USD", "USD", "USD", "H…
## $ salary_in_usd      <dbl> 79833, 260000, 109024, 20000, 150000, 72000, 190000…
## $ employee_residence <chr> "DE", "JP", "GB", "HN", "US", "US", "US", "HU", "US…
## $ remote_ratio       <dbl> 0, 0, 50, 0, 50, 100, 100, 50, 100, 50, 0, 0, 0, 10…
## $ company_location   <chr> "DE", "JP", "GB", "HN", "US", "US", "US", "HU", "US…
## $ company_size       <chr> "L", "S", "M", "S", "L", "L", "S", "L", "L", "S", "…
```

```
summary(ds$salary_in_usd)
```

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

```
ds_ft <- ds %>%
  filter(employment_type == "FT")

glimpse(ds_ft)
```

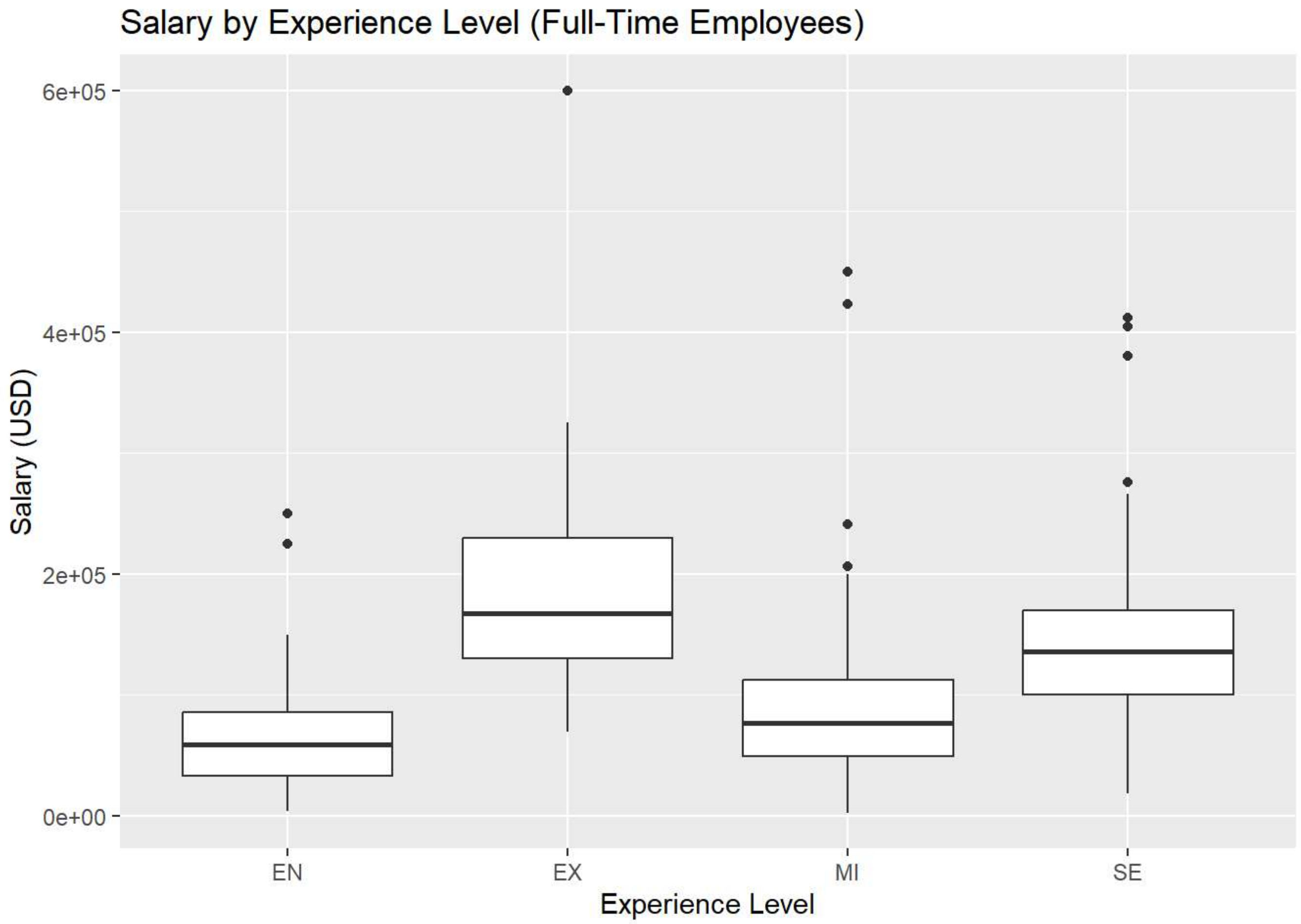
```
## Rows: 588
## Columns: 12
## $ ...1               <dbl> 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 1…
## $ work_year          <dbl> 2020, 2020, 2020, 2020, 2020, 2020, 2020, 2020, 202…
## $ experience_level   <chr> "MI", "SE", "SE", "MI", "SE", "EN", "SE", "MI", "MI…
## $ employment_type    <chr> "FT", "FT", "FT", "FT", "FT", "FT", "FT", "FT", "FT…
## $ job_title          <chr> "Data Scientist", "Machine Learning Scientist", "Bi…
## $ salary             <dbl> 70000, 260000, 85000, 20000, 150000, 72000, 190000,…
## $ salary_currency    <chr> "EUR", "USD", "GBP", "USD", "USD", "USD", "USD", "H…
## $ salary_in_usd      <dbl> 79833, 260000, 109024, 20000, 150000, 72000, 190000…
## $ employee_residence <chr> "DE", "JP", "GB", "HN", "US", "US", "US", "HU", "US…
## $ remote_ratio       <dbl> 0, 0, 50, 0, 50, 100, 100, 50, 100, 50, 0, 0, 0, 10…
## $ company_location   <chr> "DE", "JP", "GB", "HN", "US", "US", "US", "HU", "US…
## $ company_size       <chr> "L", "S", "M", "S", "L", "L", "S", "L", "L", "S", "…
```

```
exp_summary <- ds_ft %>%
  group_by(experience_level) %>%
  summarise(
    n = n(),
    mean_salary   = mean(salary_in_usd, na.rm = TRUE),
    median_salary = median(salary_in_usd, na.rm = TRUE),
    q25 = quantile(salary_in_usd, 0.25, na.rm = TRUE),
    q75 = quantile(salary_in_usd, 0.75, na.rm = TRUE)
  )

exp_summary
```

```
## # A tibble: 4 × 6
##   experience_level     n mean_salary median_salary     q25     q75
##   <chr>            <int>       <dbl>         <dbl>   <dbl>   <dbl>
## 1 EN                  79      64457.         59102  33536.  85852.
## 2 EX                  25     190728.        167875 130000  230000
## 3 MI                 206      88403.         77161  49461  112225
## 4 SE                 278     139021.        136300 100000  170000
```

```
ggplot(ds_ft, aes(x = experience_level, y = salary_in_usd)) +
  geom_boxplot() +
  labs(
    title = "Salary by Experience Level (Full-Time Employees)",
    x = "Experience Level",
    y = "Salary (USD)"
  )
```



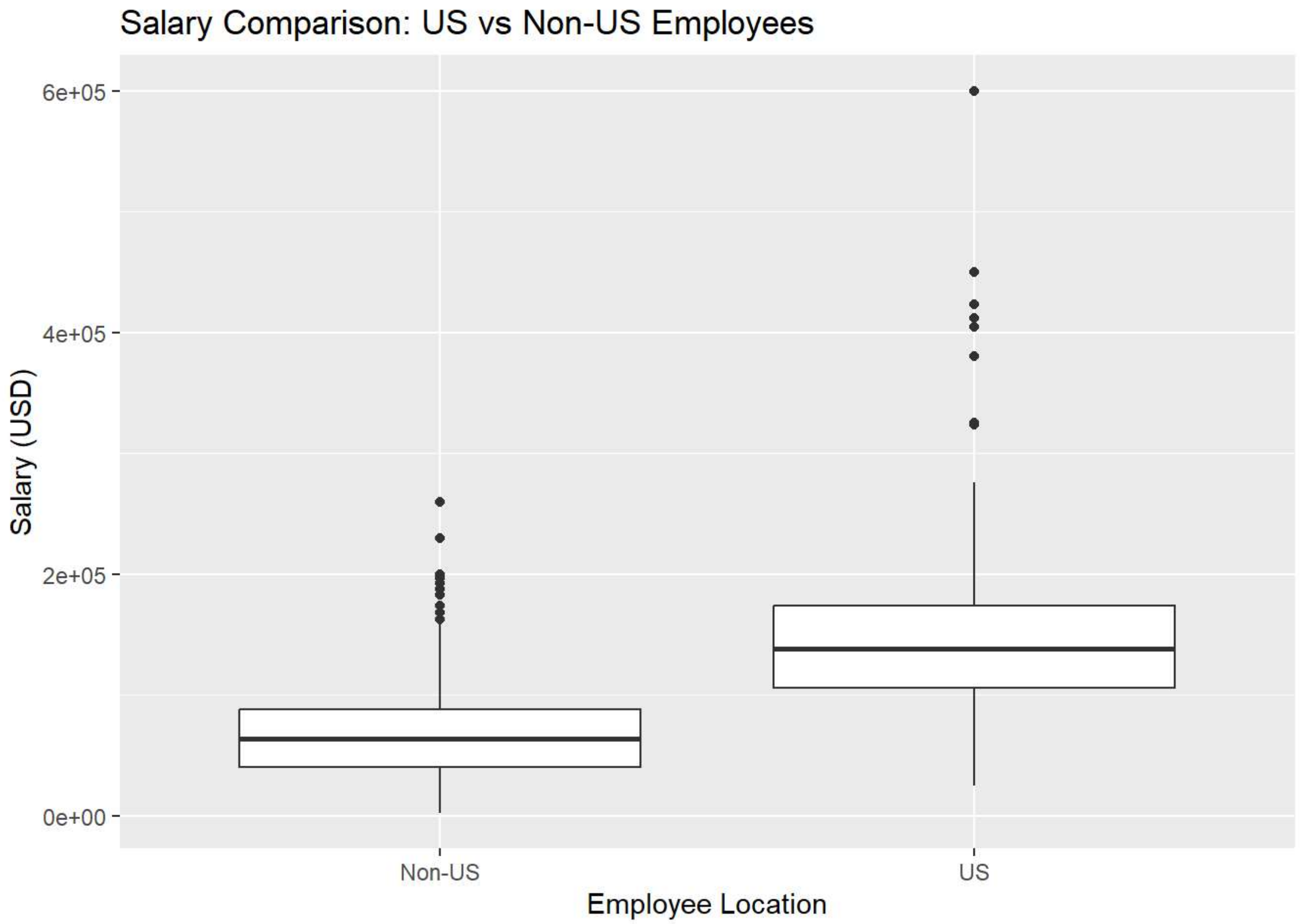
```
ds_ft <- ds_ft %>%
  mutate(
    us_employee = if_else(employee_residence == "US", "US", "Non-US")
  )

us_summary <- ds_ft %>%
  group_by(us_employee) %>%
  summarise(
    n = n(),
    mean_salary   = mean(salary_in_usd, na.rm = TRUE),
    median_salary = median(salary_in_usd, na.rm = TRUE),
    q25 = quantile(salary_in_usd, 0.25, na.rm = TRUE),
    q75 = quantile(salary_in_usd, 0.75, na.rm = TRUE)
  )

us_summary
```

```
## # A tibble: 2 × 6
##   us_employee     n mean_salary median_salary    q25    q75
##   <chr>       <int>       <dbl>         <dbl>  <dbl>  <dbl>
## 1 Non-US        260      69530.         63760.  40408  88654
## 2 US            328     148297.        138475  106195 174250
```

```
ggplot(ds_ft, aes(x = us_employee, y = salary_in_usd)) +
  geom_boxplot() +
  labs(
    title = "Salary Comparison: US vs Non-US Employees",
    x = "Employee Location",
    y = "Salary (USD)"
  )
```



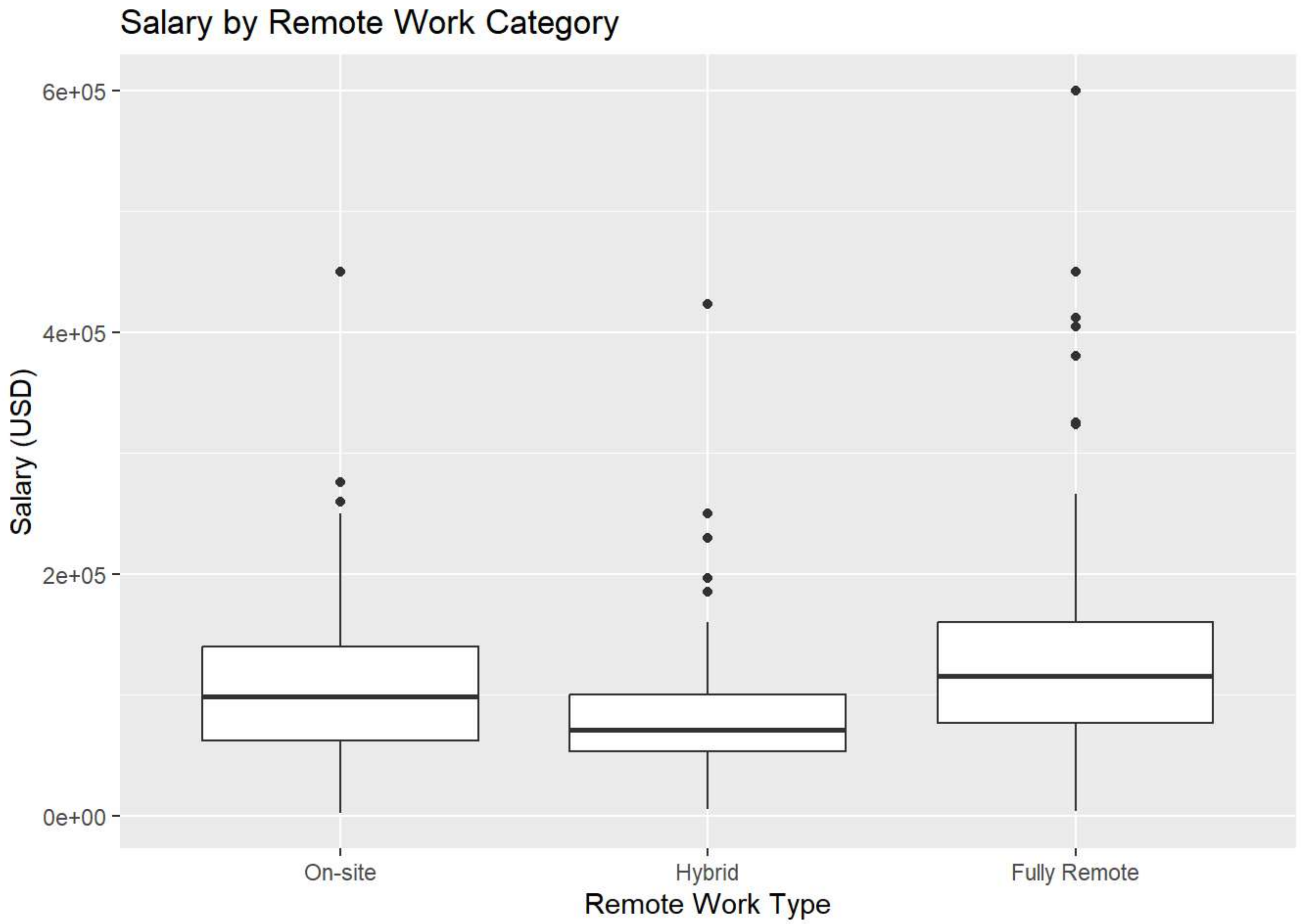
```
ds_ft <- ds_ft %>%
  mutate(
    remote_cat = factor(
      remote_ratio,
      levels = c(0, 50, 100),
      labels = c("On-site", "Hybrid", "Fully Remote")
    )
  )

remote_summary <- ds_ft %>%
  group_by(remote_cat) %>%
  summarise(
    n = n(),
    mean_salary   = mean(salary_in_usd, na.rm = TRUE),
    median_salary = median(salary_in_usd, na.rm = TRUE)
  )

remote_summary
```

```
## # A tibble: 3 × 4
##   remote_cat       n mean_salary median_salary
##   <fct>        <int>       <dbl>         <dbl>
## 1 On-site        126     107040.         99000
## 2 Hybrid          92      84440.         71562
## 3 Fully Remote   370     122875.        115717
```

```
ggplot(ds_ft, aes(x = remote_cat, y = salary_in_usd)) +
  geom_boxplot() +
  labs(
    title = "Salary by Remote Work Category",
    x = "Remote Work Type",
    y = "Salary (USD)"
  )
```



```
size_summary <- ds_ft %>%
  group_by(company_size) %>%
  summarise(
    n = n(),
    mean_salary   = mean(salary_in_usd, na.rm = TRUE),
    median_salary = median(salary_in_usd, na.rm = TRUE)
  )

size_summary
```

```
## # A tibble: 3 × 4
##   company_size     n mean_salary median_salary
##   <chr>        <int>       <dbl>         <dbl>
## 1 L              193     119665.        100800
## 2 M              318     118662.        115717
## 3 S               77      76484          69741
```

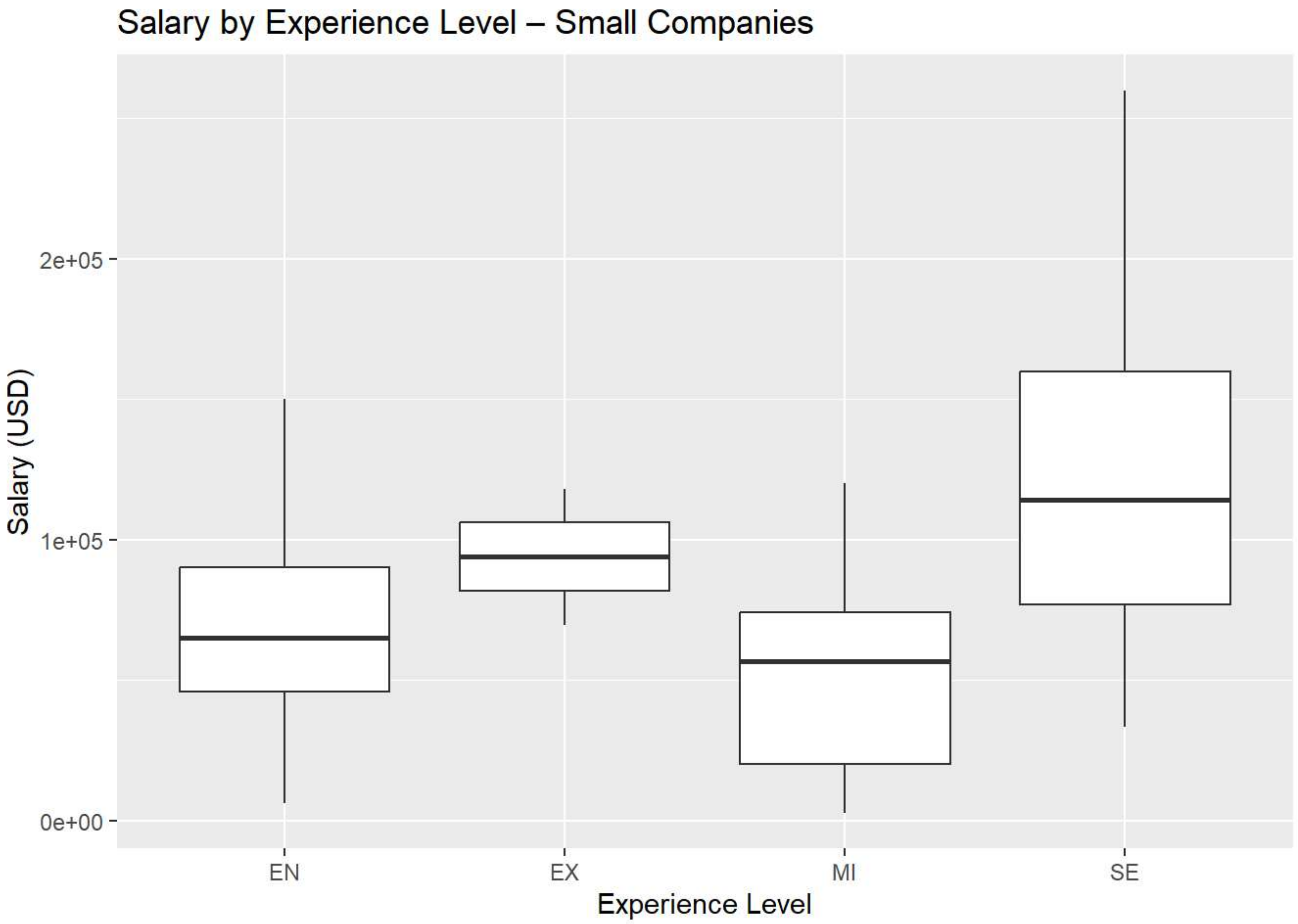
```
# Focus on small companies only
ds_small <- ds_ft %>%
  filter(company_size == "S")

small_exp_summary <- ds_small %>%
  group_by(experience_level) %>%
  summarise(
    n = n(),
    median_salary = median(salary_in_usd, na.rm = TRUE),
    q25 = quantile(salary_in_usd, 0.25, na.rm = TRUE),
    q75 = quantile(salary_in_usd, 0.75, na.rm = TRUE)
  )

small_exp_summary
```

```
## # A tibble: 4 × 5
##   experience_level     n median_salary    q25     q75
##   <chr>            <int>         <dbl>  <dbl>   <dbl>
## 1 EN                  25         65000 45896   90000
## 2 EX                   2         93964 81852. 106076.
## 3 MI                  29         56738 20000   74000
## 4 SE                  21        114047 76833  160000
```

```
ggplot(ds_small, aes(x = experience_level, y = salary_in_usd)) +
  geom_boxplot() +
  labs(
    title = "Salary by Experience Level – Small Companies",
    x = "Experience Level",
    y = "Salary (USD)"
  )
```



```
small_exp_summary
```

```
## # A tibble: 4 × 5
##   experience_level      n median_salary    q25     q75
##   <chr>             <int>         <dbl>  <dbl>   <dbl>
## 1 EN                   25         65000 45896   90000
## 2 EX                    2         93964 81852. 106076.
## 3 MI                   29         56738 20000   74000
## 4 SE                   21        114047 76833  160000
```