Simonsen.module05RProject

Steven Simonsen

2024-04-14

```
#Import dataset
r_project_data <- read.csv("~/School/DSE5002/Week_5/Project/r project data.csv")

#libraries used
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(scales)
library(ggplot2)
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

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

```
library(stringr)

#Examine summary statistics
summary(r_project_data)
```

```
##        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
```

```
head(r_project_data)
```

```
##   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
```

```
#Determine if "NA" values exist - None exist and this is good. No treatment needed.
print("Position of missing values")
```

```
## [1] "Position of missing values"
```

```
which(is.na(r_project_data))
```

```
## integer(0)
```

```
print("Count of total missing values")
```

```
## [1] "Count of total missing values"
```

```
sum(is.na(r_project_data))
```

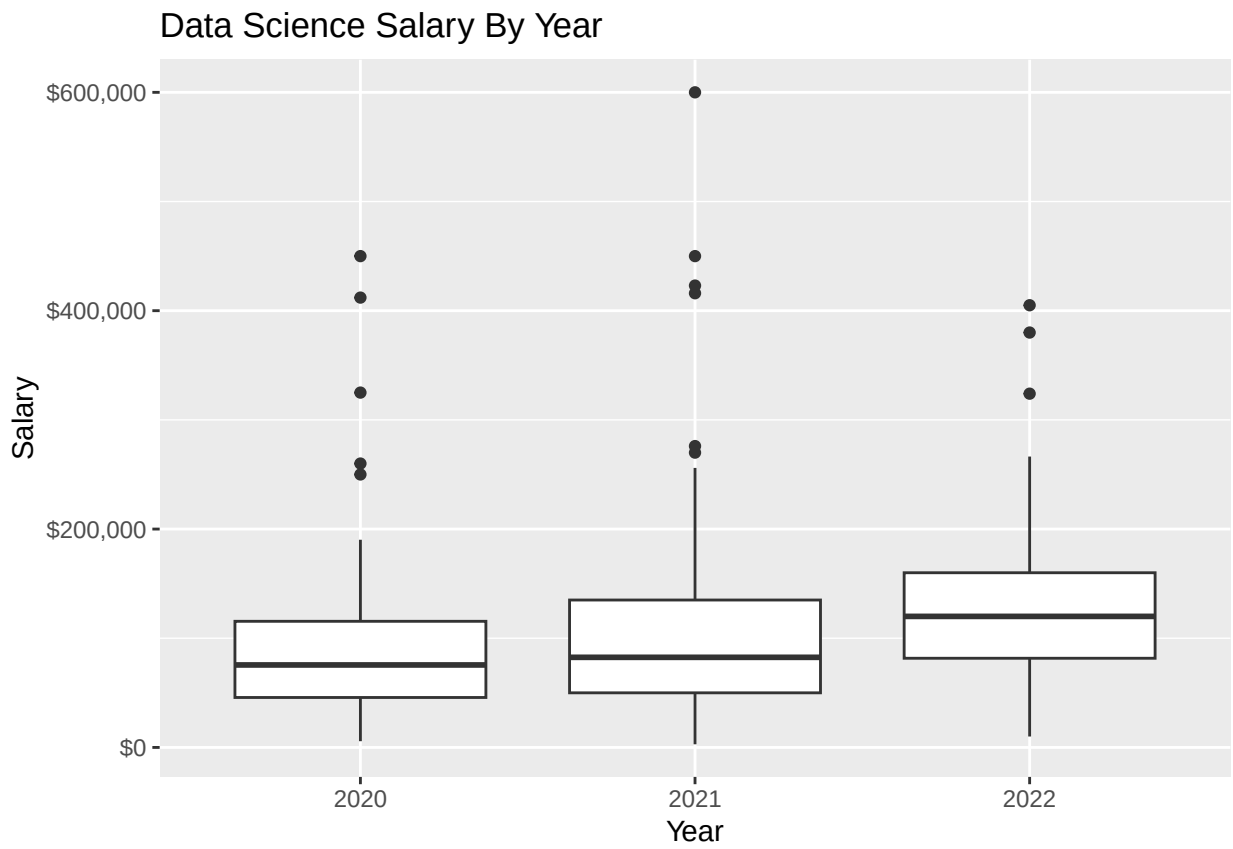
```
## [1] 0
```

```
#Initial data wrangling

#Rename column 1 to row_id
colnames(r_project_data)[1] <- c("row_id")

#Code below converts from int data type to eventual chr to better use in plots
#later
r_project_data$work_year <- ymd(paste0(r_project_data$work_year, "0101"))
```

```r
r_project_data$work_year <- format(r_project_data$work_year, "%Y")

#View project head to make sure all looks appropriate
head(r_project_data)
```

```
##   row_id 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
```

```r
#begin data aggregation and initial analysis and plots

#Box Plot of Salary
ggplot(r_project_data, aes(group=work_year)) +
  geom_boxplot(aes(x=work_year
                   ,y=salary_in_usd))+
  scale_y_continuous(labels = dollar_format(prefix = "$",
                                            big.mark = ",", decimal.mark = "."
                                            ,accuracy = 1)) +
  labs(x='Year'
       ,y='Salary'
       ,title='Data Science Salary By Year')
```

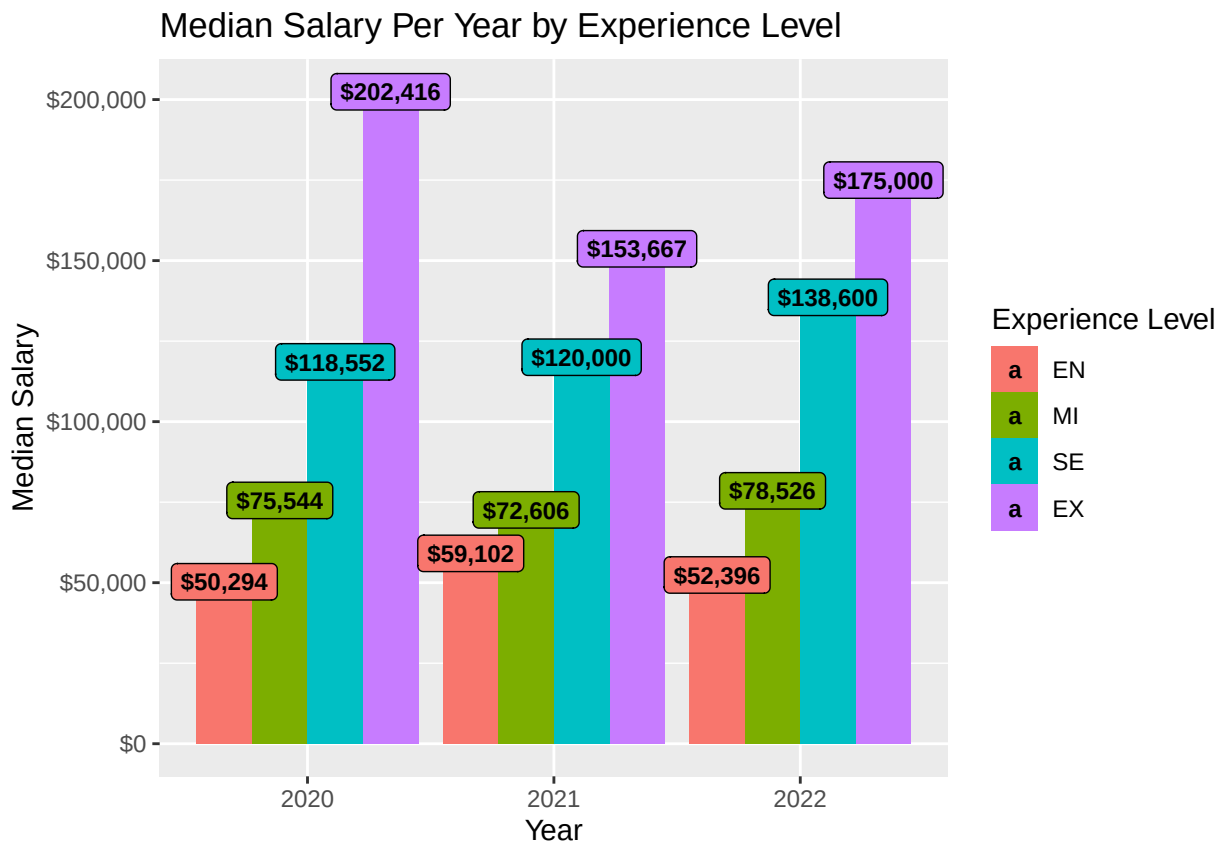



```
#Notes/Findings: Lots of outliers, median is the better measure when analyzing
#salary



#Data wrangling and Median Salary Per Year By Experience Level
#Average and median salary by work year and experience in USD
year_experience_group <- r_project_data %>%
  group_by(work_year, experience_level) %>%
  summarise(average_salary_usd=mean(salary_in_usd)
            ,median_salary_usd=median(salary_in_usd)) %>%
  arrange(experience_level)
```

```
## `summarise()` has grouped output by 'work_year'. You can override using the
## `.groups` argument.
```

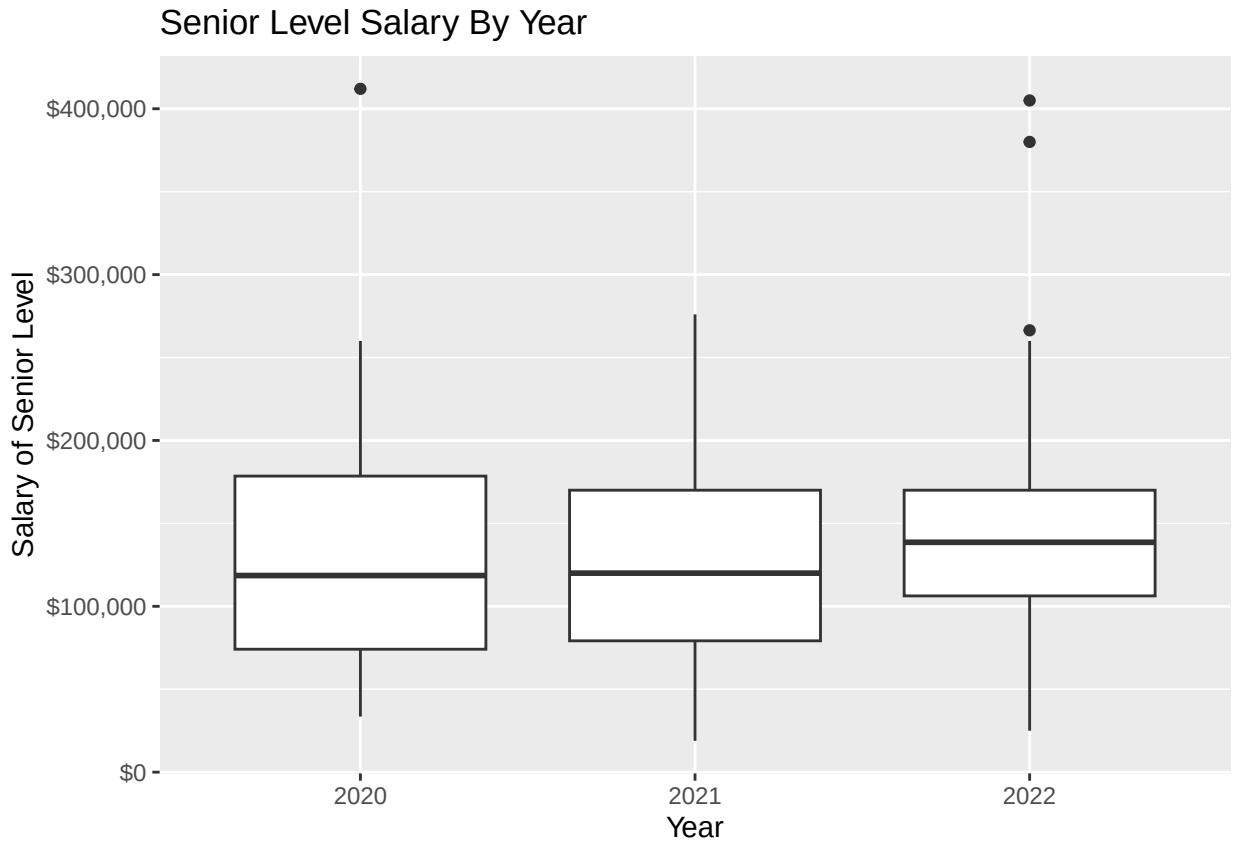
```
#Set experience level as factors to re-order for plot below
year_experience_group$experience_level <- factor(year_experience_group$experience_level
                                                 ,levels = c("EN", "MI"
                                                             ,"SE", "EX"))

#Median Salary Per Year By Experience Level
ggplot(year_experience_group, aes(x=work_year, y=median_salary_usd
                                  ,fill=experience_level)) +
  geom_col(position="dodge") +
  geom_label(aes(label=dollar(median_salary_usd, prefix = "$"
                              ,big.mark = ",", decimal.mark = "."
```

```
                              ,accuracy = 1))
            ,size=3, vjust = 0.5, position=position_dodge(0.9)
            ,fontface="bold")+
  scale_y_continuous(labels = dollar_format(prefix = "$"
                                            ,big.mark = ",", decimal.mark = "."
                                            ,accuracy = 1)) +
  labs(x='Year'
       ,y='Median Salary'
       ,fill='Experience Level'
       ,title = 'Median Salary Per Year by Experience Level')
```




```
#Notes/Findings - As expected, the highest salaries are paid with the most experience.
#However,the gap between salaries given the experience has narrowed.



#Another, closer look at Senior Level salaries boxplot
#Filter data to only SE level
salary_experience_level <- r_project_data %>%
  filter(experience_level=="SE")

#Boxplot showing Senior level salary by year
ggplot(salary_experience_level, aes(group=work_year)) +
  geom_boxplot(aes(x=work_year
                   ,y=salary_in_usd))+
  scale_y_continuous(labels = dollar_format(prefix = "$",
```

```
                                        big.mark = ",", decimal.mark = "."
                                        ,accuracy = 1)) +
  labs(x='Year'
       ,y='Salary of Senior Level'
       ,title='Senior Level Salary By Year')
```

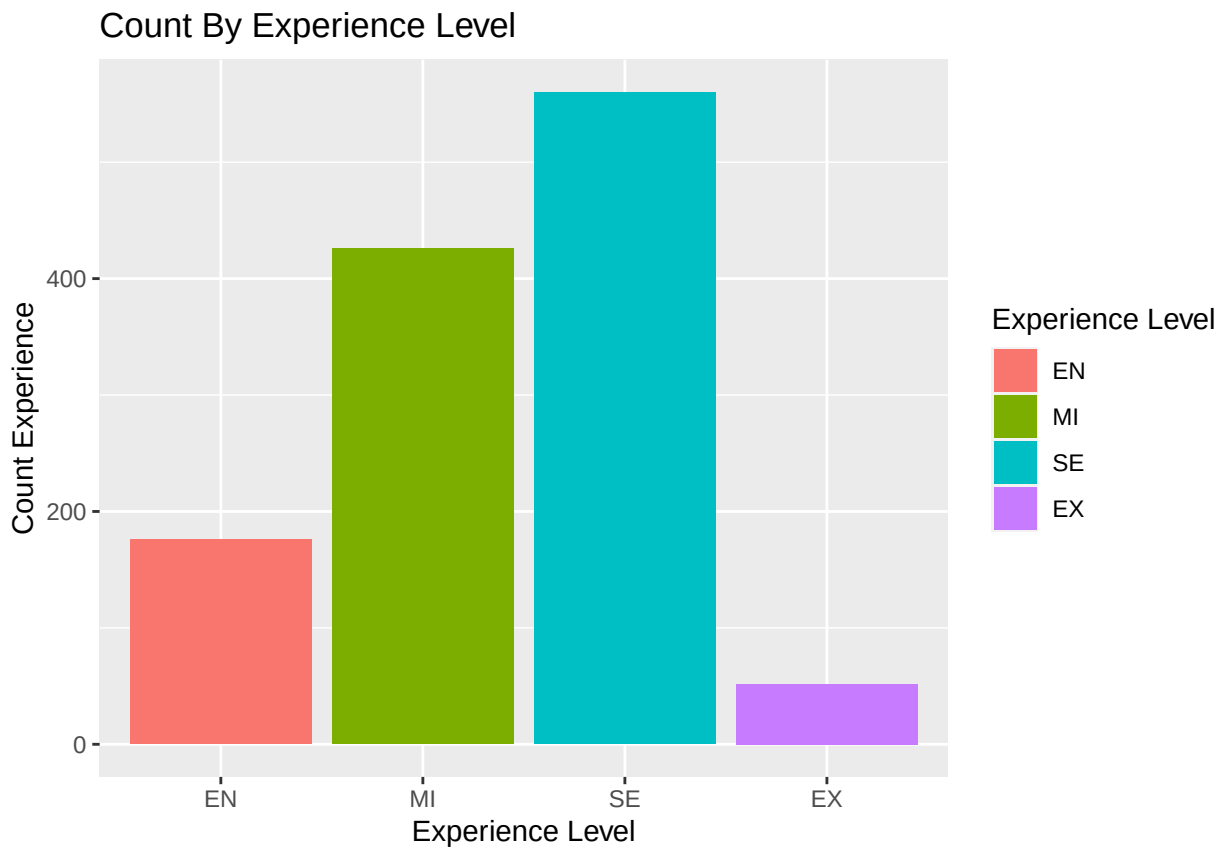



```
#Summary stats
summary(salary_experience_level)
```

```
##      row_id       work_year         experience_level   employment_type
##  Min.   :  1.0   Length:280         Length:280         Length:280
##  1st Qu.:238.2   Class :character   Class :character   Class :character
##  Median :362.5   Mode  :character   Mode  :character   Mode  :character
##  Mean   :359.2
##  3rd Qu.:516.8
##  Max.   :605.0
##    job_title             salary        salary_currency    salary_in_usd
##  Length:280         Min.   :  24000   Length:280         Min.   : 18907
##  Class :character   1st Qu.: 101378   Class :character   1st Qu.:100000
##  Mode  :character   Median : 140000   Mode  :character   Median :135500
##                     Mean   : 213949                      Mean   :138617
##                     3rd Qu.: 175025                      3rd Qu.:170000
##                     Max.   :7000000                      Max.   :412000
##  employee_residence  remote_ratio    company_location   company_size
##  Length:280         Min.   :  0.00   Length:280         Length:280
##  Class :character   1st Qu.: 50.00   Class :character   Class :character
```

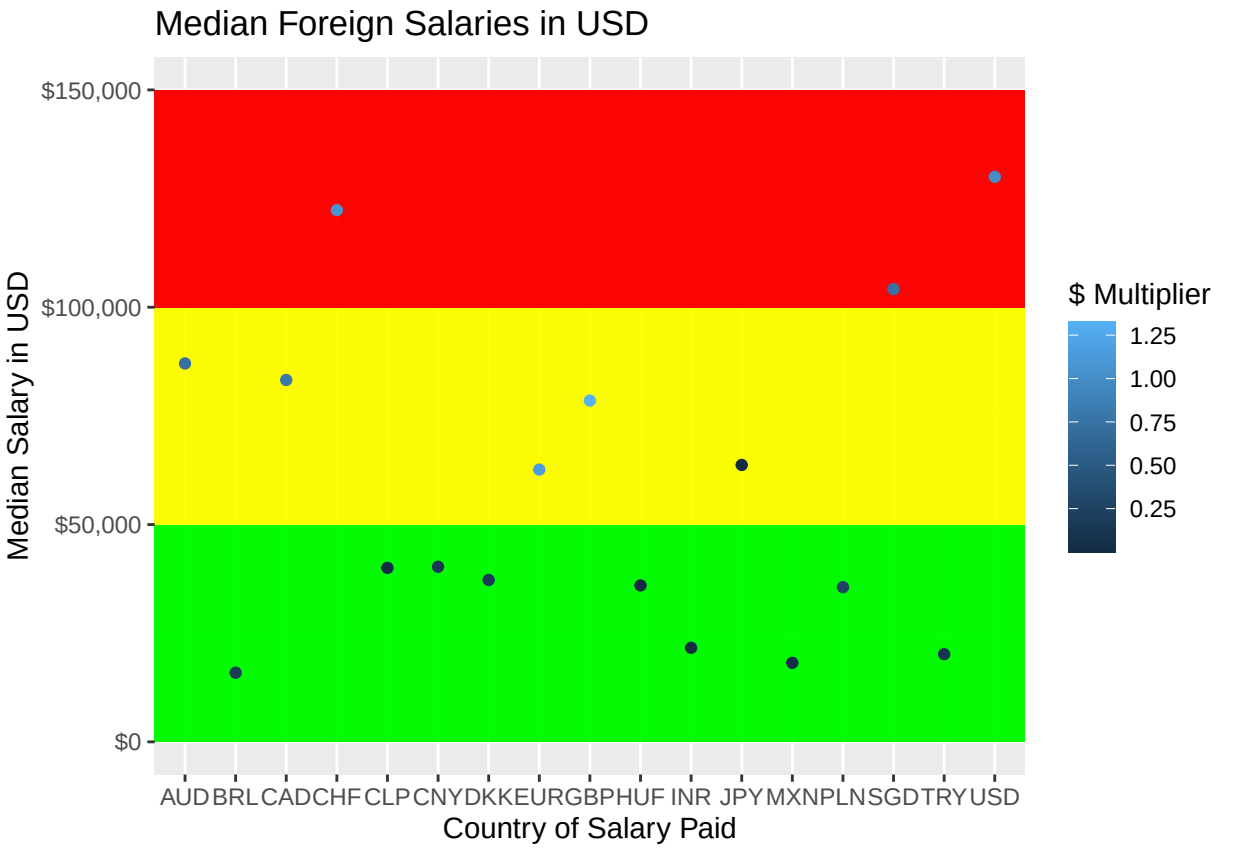
```
##  Mode  :character   Median :100.00   Mode  :character   Mode  :character
##                     Mean   : 75.89
##                     3rd Qu.:100.00
##                     Max.   :100.00
```

```r
#Notes/Findings: Some outliers in 2022, however was able to take a closer look
#At summary statistics using a boxplot.



#Count of The Data by Experience Level
#Wrangle/group data to sum experience level
count_position_experience <- r_project_data %>%
  group_by(experience_level) %>%
  reframe(count_experience=(sum(str_count(experience_level))))

#Convert experience level to factor to re-order for plot below
count_position_experience$experience_level <- factor(count_position_experience$experience_level
                                     , levels = c("EN", "MI"
                                                  ,"SE", "EX"))

#Plot Count of Experience Level
ggplot(count_position_experience) +
  geom_col(aes(x=experience_level,y=count_experience
               ,fill=experience_level))+
  labs(x='Experience Level'
       ,y='Count Experience'
       ,fill='Experience Level'
       ,title='Count By Experience Level')
```




```
#Notes/Findings: The most experience exists in in the MI and SE level within
#this dataset.



#Median Salary in USD and exchange rates for International Analysis
#Exchange Rate Equation - Multiply salary by multiplier_to_usd to get
#salary in USD
currency_exchange_rate <- r_project_data %>%
  group_by(salary_currency) %>%
  summarise(multiplier_to_usd=mean(salary_in_usd/salary))

#Median salary in usd given the country
median_salary_international <- r_project_data %>%
  group_by(salary_currency) %>%
  summarise(median_salary_usd=median(salary_in_usd))

#Join tables together to get exchange rate multiplier and median salary in usd
joined_salary_exchange <- currency_exchange_rate %>%
  inner_join(median_salary_international, by='salary_currency')

#Plot median salary and color by exchange rate multiplier
ggplot(joined_salary_exchange) +
  geom_rect(aes(xmin=-Inf, xmax=Inf, ymin=0, ymax=50000), fill="green", alpha=0.2)+
  geom_rect(aes(xmin=-Inf, xmax=Inf, ymin=50000, ymax=100000), fill="yellow", alpha=0.2)+
  geom_rect(aes(xmin=-Inf, xmax=Inf, ymin=100000, ymax=150000), fill="red", alpha=0.2)+
```

```
geom_point(aes(x=salary_currency, y=median_salary_usd
               ,color=multiplier_to_usd))+
scale_y_continuous(labels = dollar_format(prefix = "$",
                                          big.mark = ",", decimal.mark = "."
                                          ,accuracy = 1)) +
labs(x='Country of Salary Paid'
     ,y='Median Salary in USD'
     ,color='$ Multiplier'
     ,title='Median Foreign Salaries in USD')
```




```
#Notes/Findings: Green means a lower median salary given the country to usd,
#and red is the highest median salary in usd. The gradient of the dots show
#how low or high the exchange rate is in relation to the salary paid.
#So, a darker dot means the salary is worth less in relation to usd.
```