Visualization Activity

Melissa Jayawardane

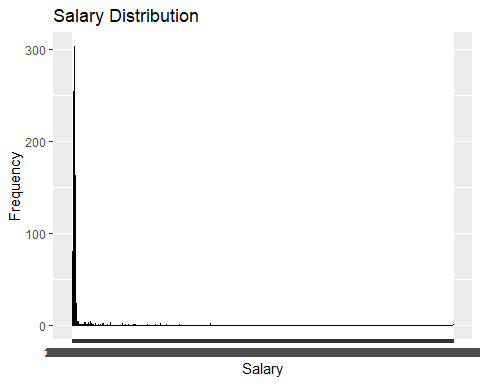
2023-06-19

## Data Visualization

# Read the dataset into R using read.csv() function  
df<-read.csv("C:/Users/user/Documents/ds\_salaries.csv")   
str(df)

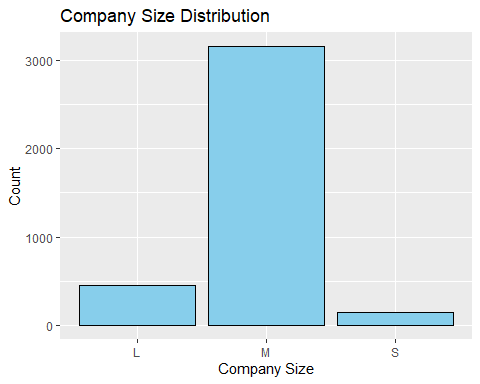
## 'data.frame': 3755 obs. of 11 variables:  
## $ work\_year : int 2023 2023 2023 2023 2023 2023 2023 2023 2023 2023 ...  
## $ experience\_level : chr "SE" "MI" "MI" "SE" ...  
## $ employment\_type : chr "FT" "CT" "CT" "FT" ...  
## $ job\_title : chr "Principal Data Scientist" "ML Engineer" "ML Engineer" "Data Scientist" ...  
## $ salary : int 80000 30000 25500 175000 120000 222200 136000 219000 141000 147100 ...  
## $ salary\_currency : chr "EUR" "USD" "USD" "USD" ...  
## $ salary\_in\_usd : int 85847 30000 25500 175000 120000 222200 136000 219000 141000 147100 ...  
## $ employee\_residence: chr "ES" "US" "US" "CA" ...  
## $ remote\_ratio : int 100 100 100 100 100 0 0 0 0 0 ...  
## $ company\_location : chr "ES" "US" "US" "CA" ...  
## $ company\_size : chr "L" "S" "S" "M" ...

library(ggplot2)  
# Calculate the range of the data  
data\_range <- max(df$salary) - min(df$salary)  
  
# Estimate the bin width using the Freedman-Diaconis rule  
bin\_width <- 2 \* IQR(df$salary) / (length(df$salary)^(1/3))  
  
# Calculate the number of groups  
num\_groups <- ceiling(data\_range / bin\_width)  
  
# Create a histogram with the estimated number of groups  
ggplot(df, aes(x = salary)) +  
 geom\_histogram(binwidth = bin\_width, fill = "skyblue", color = "black") +  
 labs(title = "Salary Distribution",  
 x = "Salary",  
 y = "Frequency") +  
 scale\_x\_continuous(breaks = seq(min(df$salary), max(df$salary), by = bin\_width))



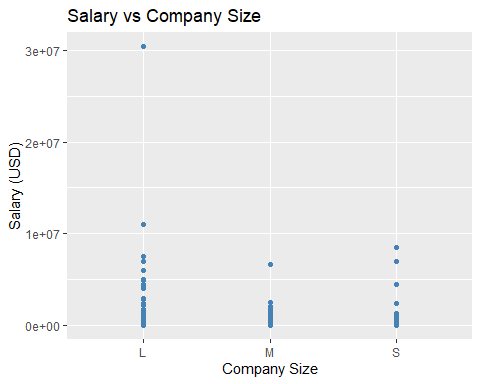
The above histogram represents the salary distribution across the market. It is evident from the plot that the distribution is skewed to the left, with majority of the employees in the industry having lower-end salaries.

# Create a bar plot of company size distribution  
ggplot(df, aes(x = company\_size)) +  
 geom\_bar(fill = "skyblue", color = "black") +  
 labs(title = "Company Size Distribution",  
 x = "Company Size",  
 y = "Count")



Based on the above plot, most of the companies are medium size. If the majority of companies in the dataset are categorized as medium size, it implies that there is a prevalence of companies falling within this size range. This finding can lead to several interpretations and inferences. The dominance of medium-sized companies may indicate a balanced market landscape with a mix of both small and large competitors.

# Scatterplot of Salary and Company Size  
ggplot(df, aes(x = company\_size, y = salary)) +  
 geom\_point(color = "steelblue") +  
 labs(title = "Salary vs Company Size", x = "Company Size", y = "Salary (USD)")



The scatterplot indicates that the medium sized companies have a lower salary range compared to small and large companies. The salary margins between employees is significantly lower when compared to other companies.