ANA 515 Assignment 2

Melissa Jayawardane

2023-06-13

## Section 1

The gathered data in this dataset provides tremendous value to those seeking knowledge regarding job salaries within the domain of data science. The compiled information includes vast amounts of detail about various facets of salary payment across sectors like work year, experience level and employment type - all critical indicators of earning potential. Any potential user could analyze this comprehensive table’s rows and columns - 11 columns to be exact - which breaks down everything from job title to remote work allowances amongst various companies across countless industries. Of note, this abundant repository’s CSV file format affords those interested ease-of-use flexibility -the ability to hash-out relevant figures- via specialized software packages available in several modern programming languages like R.

## Section 2

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

## Section 3

# Load the required packages  
library(tidyverse)  
library(knitr)  
library(dplyr)  
library(kableExtra)  
# Rename columns  
df <- df %>%  
 rename(year = work\_year,  
 experience\_level = experience\_level,  
 employment = employment\_type,  
 job\_title = job\_title,  
 salary = salary,  
 currency = salary\_currency,  
 salary\_usd = salary\_in\_usd,  
 employee\_residence = employee\_residence,  
 remote\_ratio = remote\_ratio,  
 company\_location = company\_location,  
 company\_size = company\_size)  
  
# Subset to keep at least 4 columns  
sub\_df <- df %>%  
 select(year, experience\_level, job\_title, salary, remote\_ratio, salary\_usd)

## Section 4

This dataframe has 3755 rows and 6 columns. The names of the columns and a brief description of each are in the table below

# Define the column names and descriptions as a data frame  
column\_description <- data.frame(Column = colnames(sub\_df), Description = c("The year the salary was paid",  
 "The experience level in the job",  
 "The job title",  
 "The total gross salary amount",  
 "The overall amount of work done remotely",  
 "The salary in USD"))  
# Display the table  
knitr::kable(column\_description, format = "markdown", col.names = c("Column", "Description"))

| Column | Description |
| --- | --- |
| year | The year the salary was paid |
| experience\_level | The experience level in the job |
| job\_title | The job title |
| salary | The total gross salary amount |
| remote\_ratio | The overall amount of work done remotely |
| salary\_usd | The salary in USD |

## Section 5

# Select three columns  
selected\_columns <- sub\_df[, c("remote\_ratio", "salary\_usd", "salary")]  
names(sub\_df)

## [1] "year" "experience\_level" "job\_title" "salary"   
## [5] "remote\_ratio" "salary\_usd"

# Calculate summaries for each column  
summary\_data <- summary(selected\_columns)  
min\_values <- sapply(selected\_columns, min, na.rm = TRUE)  
max\_values <- sapply(selected\_columns, max, na.rm = TRUE)  
mean\_values <- sapply(selected\_columns, mean, na.rm = TRUE)  
num\_missing <- sapply(selected\_columns, function(x) sum(is.na(x)))  
  
# Assign the summaries to a new object  
column\_summaries <- data.frame(Column = colnames(selected\_columns),  
 Minimum = min\_values,  
 Maximum = max\_values,  
 Mean = mean\_values,  
 "Number of Missing Values" = num\_missing)  
  
# Display the column summaries  
column\_summaries

## Column Minimum Maximum Mean  
## remote\_ratio remote\_ratio 0 100 46.27164  
## salary\_usd salary\_usd 5132 450000 137570.38988  
## salary salary 6000 30400000 190695.57177  
## Number.of.Missing.Values  
## remote\_ratio 0  
## salary\_usd 0  
## salary 0