Whether Sexism exists in the IT industry?

Yang Song

May 03, 2022

# Installing ‘rmarkdown’ Package

install.packages("rmarkdown")  
install.packages("knitr")  
library(rmarkdown)  
library(knitr)

##Seting global options

# Reading Excel files

## Save the csv file into my working directory as an Excel sheet first

## [1] "C:/Users/songy/Desktop/663-R/Project-663/My-R-Project"

## Loading required package: plyr

## Warning: package 'plyr' was built under R version 4.1.3

## Loading required package: dplyr

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:plyr':  
##   
## arrange, count, desc, failwith, id, mutate, rename, summarise,  
## summarize

## The following objects are masked from 'package:stats':  
##   
## filter, lag

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

# selecting certain variables to study

## Remove all the rows with NA

D2 <- na.omit(D1)

## Cleaning the data & Specifying certain rows to study

a <- which(D2$CompTotal== 'Monthly')  
D2$CompTotal[a] <- D2$CompTotal[a]\*12  
b <- which(D2$CompTotal== 'Weekly')  
D2$CompTotal[b] <- D2$CompTotal[b]\*52  
  
c <- which(D2$YearsCodePro== 'Less than 1 year')  
D2$YearsCodePro[c] <- 0.5  
d <- which(D2$YearsCodePro=='More than 50 years')  
D2$YearsCodePro[d] <- 55  
D2$YearsCodePro<-as.numeric(D2$YearsCodePro)  
  
Gender\_Female <- which(D2$Gender == 'Woman')  
D2$Gender[Gender\_Female] <- 1  
Gender\_Male <- which(D2$Gender == 'Man')  
D2$Gender[Gender\_Male] <- 0  
D2$Gender <- as.numeric(D2$Gender)

## Warning: NAs introduced by coercion

e<-which(D2$Age=='18-24 years old')  
D2$Age[e]<-1  
f<-which(D2$Age=='25-34 years old')  
D2$Age[f]<-2  
g<-which(D2$Age=='35-44 years old')  
D2$Age[g]<-3  
h<-which(D2$Age=='45-54 years old')  
D2$Age[h]<-4  
i<-which(D2$Age=='55-64 years old')  
D2$Age[i]<-5  
j<-which(D2$Age=='65 years or older')  
D2$Age[j]<-6  
k<-which(D2$Age=='Prefer not to say')  
D2$Age[k]<-7  
l<-which(D2$Age=='Under 18 years old')  
D2$Age[l]<-0  
D2$Age <- as.numeric(D2$Age)  
  
m <- which(D2$YearsCode== 'Less than 1 year')  
D2$YearsCode[m] <- 0.5  
n <- which(D2$YearsCode=='More than 50 years')  
D2$YearsCodePro[n] <- 55  
D2$YearsCode<-as.numeric(D2$YearsCode)

## Warning: NAs introduced by coercion

D3 <- D2 %>% select (CompTotal,Gender, Age,YearsCodePro,YearsCode)  
  
install.packages("ggplot2")

## Installing package into 'C:/Users/songy/Documents/R/win-library/4.1'  
## (as 'lib' is unspecified)

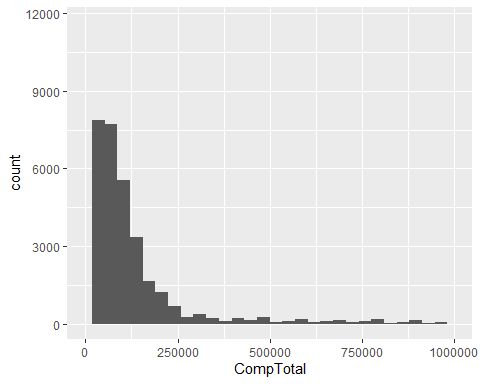
## Error in contrib.url(repos, "source"): trying to use CRAN without setting a mirror

library("ggplot2")  
ggplot(data=D2)+geom\_histogram(aes(x=CompTotal))+xlim(0,1000000)

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 3343 rows containing non-finite values (stat\_bin).

## Warning: Removed 2 rows containing missing values (geom\_bar).



D3 <- D2 %>% filter(CompTotal%in% c(0,250000))

# Insert a table D3

## Warning: package 'knitr' was built under R version 4.1.3

***2021 Survey Results With Studied Variables***

| Currency | CompTotal | CompFreq | Gender | Age | YearsCodePro | YearsCode |
| --- | --- | --- | --- | --- | --- | --- |
| RSD Serbian dinar | 250000 | Monthly | 0 | 2 | 5 | 7 |
| PLN Polish zloty | 0 | Weekly | 0 | 0 | 1 | 2 |
| INR Indian rupee | 0 | Yearly | 0 | 1 | 5 | 11 |
| USD United States dollar | 0 | Yearly | 0 | 4 | 15 | 40 |
| USD United States dollar | 250000 | Yearly | 0 | 4 | 31 | 50 |
| USD United States dollar | 250000 | Yearly | 0 | 1 | 8 | 11 |

#Summary of D3

summary(D3)

## Currency CompTotal CompFreq Gender   
## Length:441 Min. : 0 Length:441 Min. :0.00000   
## Class :character 1st Qu.: 0 Class :character 1st Qu.:0.00000   
## Mode :character Median :250000 Mode :character Median :0.00000   
## Mean :178571 Mean :0.03066   
## 3rd Qu.:250000 3rd Qu.:0.00000   
## Max. :250000 Max. :1.00000   
## NA's :17   
## Age YearsCodePro YearsCode   
## Min. :0.000 Min. : 0.50 Min. : 0.50   
## 1st Qu.:2.000 1st Qu.: 5.00 1st Qu.: 9.00   
## Median :2.000 Median :10.00 Median :16.00   
## Mean :2.646 Mean :12.64 Mean :18.70   
## 3rd Qu.:3.000 3rd Qu.:20.00 3rd Qu.:25.75   
## Max. :7.000 Max. :55.00 Max. :50.00   
## NA's :3

# Multiple Regression

## 'data.frame': 441 obs. of 7 variables:  
## $ Currency : chr "RSD\tSerbian dinar" "PLN\tPolish zloty" "INR\tIndian rupee" "USD\tUnited States dollar" ...  
## $ CompTotal : num 250000 0 0 0 250000 250000 250000 0 0 0 ...  
## $ CompFreq : chr "Monthly" "Weekly" "Yearly" "Yearly" ...  
## $ Gender : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ Age : num 2 0 1 4 4 1 2 2 4 2 ...  
## $ YearsCodePro: num 5 1 5 15 31 8 10 0.5 25 3 ...  
## $ YearsCode : num 7 2 11 40 50 11 15 5 44 7 ...  
## - attr(\*, "na.action")= 'omit' Named int [1:37133] 1 2 3 4 5 6 7 8 9 11 ...  
## ..- attr(\*, "names")= chr [1:37133] "1" "2" "3" "4" ...

##   
## Call:  
## lm(formula = CompTotal ~ Gender + Age + YearsCodePro, data = D3,   
## subset = YearsCode)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -239499 -108993 16262 99683 142890   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 116749 14098 8.281 1.83e-15 \*\*\*  
## Gender 126106 80594 1.565 0.118   
## Age -12519 9948 -1.258 0.209   
## YearsCodePro 5761 1155 4.989 9.05e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 113600 on 402 degrees of freedom  
## (34 observations deleted due to missingness)  
## Multiple R-squared: 0.1385, Adjusted R-squared: 0.1321   
## F-statistic: 21.55 on 3 and 402 DF, p-value: 5.832e-13