script

```r
# load dataset that has been cropped for analysis
data <- read.csv("Cropped_Data.csv")

# the first column of the dataset is the names of the rows
# so we renamed the row names based on first column and
#deleted the column at the end
row_number <- nrow(data)
for (i in 1:row_number) {
  rownames(data)[i] = data[i,1]
}

# Delete the column
data <- data[c(2:ncol(data))]

# Get rid of the first letter X for each of the years
colnames(data) <- gsub("X", "", colnames(data))

# Rename part of the row names
#(i.e. change .i Inapplicable into Inapplicable for better
#data representation)
rownames(data)[1] <- "Inapplicable"
rownames(data)[2] <- "No answer"
rownames(data)[3] <- "Do not Know/Cannot Choose"
rownames(data)[4] <- "Skipped on Web"

# Save the cleaned data
write.csv(data,"Cleaned_Data.csv")
```

Average Working Hour in 2008, 2021, 2022, and total

```
# load dataset that has been cropped for analysis
data <- read.csv("Cropped_Data.csv")

# the first column of the dataset is the names of the rows
# so we renamed the row names based on first column and
#deleted the column at the end
row_number <- nrow(data)
for (i in 1:row_number) {
  rownames(data)[i] = data[i,1]
}

# Delete the column
data <- data[c(2:ncol(data))]

# Get rid of the first letter X for each of the years
colnames(data) <- gsub("X", "", colnames(data))

# Rename part of the row names
#(i.e. change .i Inapplicable into Inapplicable for better
#data representation)
rownames(data)[1] <- "Inapplicable"
rownames(data)[2] <- "No answer"
rownames(data)[3] <- "Do not Know/Cannot Choose"
rownames(data)[4] <- "Skipped on Web"

# Save the cleaned data
write.csv(data,"Cleaned_Data.csv")


#Graphs
library(ggplot2)
data <- read.csv("Cleaned_Data.csv")
# Get rid of the first letter X for each of the years
colnames(data)[1] <- "work_hours"
colnames(data) <- gsub("X", "", colnames(data))

#Histogram of average working hours for 2008, 2021, 2022, and Total
hist_data <- select(data, work_hours,"2008","2021","2022","Total")
hist_data <- hist_data[5:94,]
```

```
sum <- c(0,0,0,0)
total_people <- c(0,0,0,0)
for(i in 1:4){
  total_people[i] <- sum(hist_data[, i+1])
}
hist_data[, 1] <- sapply(hist_data[, 1], as.numeric)
```

```
Warning in lapply(X = X, FUN = FUN, ...): NAs introduced by coercion
```

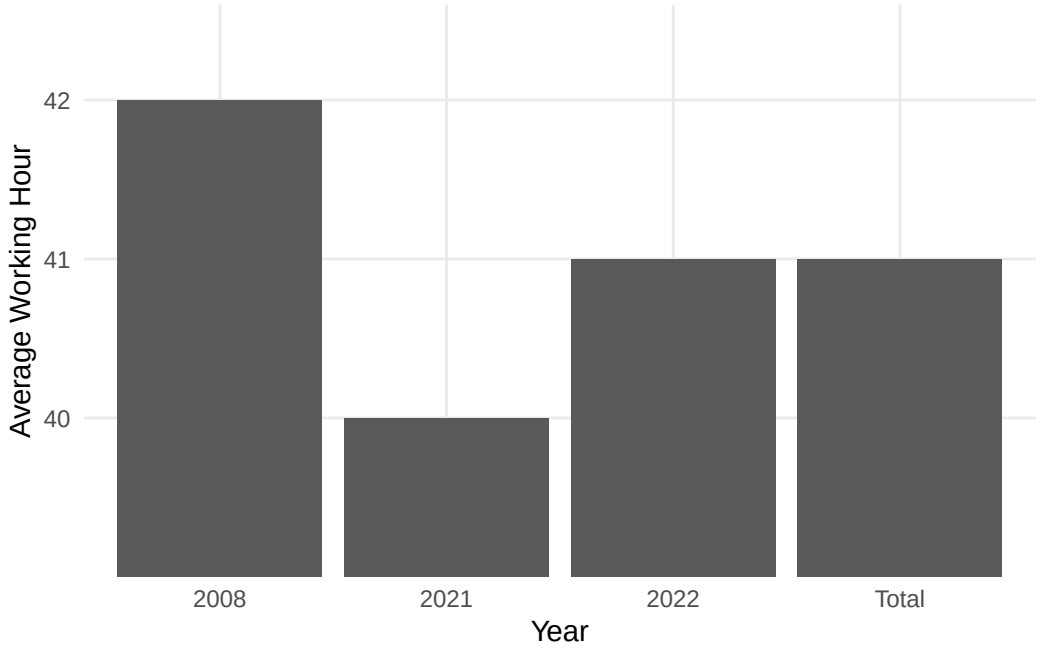
```
hist_data[90,1] <- 90

sum[1] <- sum(hist_data$"2008" * hist_data$work_hours)
sum[2] <- sum(hist_data$"2021" * hist_data$work_hours)
sum[3] <- sum(hist_data$"2022" * hist_data$work_hours)
sum[4] <- sum(hist_data$Total * hist_data$work_hours)

averages <- round(sum/total_people)
years <- c("2008","2021","2022" ,"Total")

average_hours <- data.frame(cbind(averages,years))

ggplot(average_hours,aes(x=years,y=averages)) +
  geom_bar(stat="identity") +
  theme_minimal() + # Make the theme neater
  labs(x = "Year", y = "Average Working Hour") +
  scale_color_brewer(palette = "Set1") +
  theme(legend.position = "bottom")
```



Data Cleaning: Modify hours into 1-20,20-40,40-60,60-80, 80+ categories

```
data <- read.csv("Cleaned_Data.csv")
colnames(data)[1] <- "work_hours"
colnames(data) <- gsub("X", "", colnames(data))

cate_data <- data
year <- colnames(data)
Hours <- c("No Response","0-20","21-40","41-60","61-80","80+","Total")

col_number <- ncol(cate_data) -1

filtered_df1 <- cate_data %>%
  filter(work_hours < 20)

tweenties <- rep(0, 35)
for(i in 1:col_number){
  tweenties[i] <- sum(filtered_df1 [, i+1])
}

filtered_df2 <- cate_data %>%
```

```
  filter(work_hours < 40) %>%
  filter(work_hours > 20)

forties <- rep(0, 35)
for(i in 1:col_number){
  forties[i] <- sum(filtered_df2[, i+1])
}

filtered_df3 <- cate_data %>%
  filter(work_hours < 60) %>%
  filter(work_hours > 40)

sixties <- rep(0, 35)
for(i in 1:col_number){
  sixties[i] <- sum(filtered_df3 [, i+1])
}

filtered_df4 <- cate_data %>%
  filter(work_hours < 80) %>%
  filter(work_hours > 60)

eighties <- rep(0, 35)
for(i in 1:col_number){
  eighties[i] <- sum(filtered_df4[, i+1])
}

filtered_df5 <- cate_data %>%
  filter(work_hours > 80)
filtered_df5 <- filtered_df5[5:14,]

more <- rep(0, 35)
for(i in 1:col_number){
  more[i] <- sum(filtered_df5[, i+1])
}

filtered_df6 <- cate_data %>%
  filter(work_hours > 80)
filtered_df6 <- filtered_df6[1:4,]

No_Response <- rep(0, 35)
for(i in 1:col_number){
  No_Response[i] <- sum(filtered_df6[, i+1])
```

```
}

rm(filtered_df1,filtered_df2,filtered_df3,filtered_df4,filtered_df5,filtered_df6)

total <- data[95,2:36]

cate_data <- rbind(No_Response,tweenties,forties,sixties,eighties,more,total)
cate_data <- data.frame(cbind(Hours,cate_data))
colnames(cate_data) <- year

write.csv(cate_data,"cleaned_categorized_data.csv")
```

```
data<- read.csv("cleaned_categorized_data.csv")
# Delete the column
data <- data[c(2:ncol(data))]
colnames(data)[1] <- "Work Hours/Years"
# Get rid of the first letter X for each of the years
colnames(data) <- gsub("X", "", colnames(data))
kable(data,row.names = FALSE)|>
  kable_styling() |>
  row_spec(6, hline_after = TRUE)
```

Work Hours/Years	1972	1973	1974	1975	1976	1977	1978	1980	1982	1983	1984	1985	198
No Response	1613	715	724	695	731	644	658	627	817	681	580	604	60
0-20	0	53	46	52	47	43	55	44	60	50	49	38	3
21-40	0	167	126	175	151	154	162	149	231	211	167	175	16
41-60	0	167	177	181	155	212	211	196	222	200	224	237	22
61-80	0	34	22	25	31	22	43	45	42	45	46	44	5
80+	0	8	5	9	9	6	18	13	13	14	15	20	1
Total	1613	1503	1485	1492	1496	1520	1541	1478	1869	1608	1476	1537	148

Comparsion between the years around 2008

Percentage of Non-response rate

```
data <- read.csv("cleaned_categorized_data.csv")
data <- data[c(2:ncol(data))]
```