```
---
title: "Final_Report"
output: html_document
date: "2025-11-24"
---
# Day 1 Task 2

#--- Upload the provided NHIS 2021 dataset

View(NHIS_Data_2021)

#--- load the necessary libraries

library(tidyverse)

library(ggplot2)

library(haven)

library(modeest)

library(psych)


#--- Perform an initial exploration: use str(), summary(), and head()

str(NHIS_Data_2021)

summary(NHIS_Data_2021)

head(NHIS_Data_2021)




# Day 2 Tasks 1 & 2

# Load library data for visualization and wrangling
library(tidyverse)

# Load data to be stored within the current file
setwd("~/Downloads") # Data will be stored here
getwd() # Working directory
list.files() # List files
nhis <- read_csv("NHIS _Data_2021.csv") # NHIS dataset

# Check data
glimpse(nhis) # Quick look at data
summary(nhis) # Description of variables

# Sort variables for analysis
nhis_selected <- nhis %>%
  select(AGEP_A, WEIGHTLBTC_A, HEIGHTTC_A, SEX_A, HISPALLP_A, EDUCP_A, PHSTAT_A,
LSATIS4R_A)

# Define missing or invalid code
missing_codes <- c(7, 8, 9, 97, 98, 99, 996, 997, 998, 999)

# Check data set variables and environment
ls()
names(nhis)

# Clean data
nhis_selected <- nhis %>%
  select(
    AGEP_A, WEIGHTLBTC_A, HEIGHTTC_A, SEX_A, HISPALLP_A, EDUCP_A, PHSTAT_A, LSATIS4R_A
  )
head(nhis_selected)

# Remove unnecessary data
nhis_clean_task1 <- nhis_selected %>%
  filter(!(
    AGEP_A %in% missing_codes |
      WEIGHTLBTC_A %in% missing_codes |
      HEIGHTTC_A %in% missing_codes |
      SEX_A %in% missing_codes |
      HISPALLP_A %in% missing_codes |
      EDUCP_A %in% missing_codes |
      PHSTAT_A %in% missing_codes |
      LSATIS4R_A %in% missing_codes
  ))

# Check rows
nrow(nhis_selected)
nrow(nhis_clean_task1)

# Recode variables
nhis_clean <- nhis_clean_task1 %>%
  mutate(
    EDUCP_A_recoded = case_when(
      EDUCP_A %in% 0:3 ~"Less than High School",
      EDUCP_A == 4 ~ "High School Graduate",
      EDUCP_A %in% 5:7 ~ "Some College Education",
      EDUCP_A %in% 8:10 ~ "College Graduate or better",
      TRUE ~ NA_character_
    )
  )

# Convert the recoded data in order
nhis_clean$EDUCP_A_recoded <- factor(
  nhis_clean$EDUCP_A_recoded,
  levels = c(
    "Less than High School",
    "High School Graduate",
    "Some College Education",
    "College Graduate or better"
  )
)

# Check frequency of education
table(nhis_clean$EDUCP_A_recoded)

# Create folder and clean data
dir.create("data", showWarnings = FALSE)
write_csv(nhis_clean, "data/nhis_clean.csv")

# Save content
list.files("data")

# Save a copy
write.csv(nhis_clean, "nhis_clean.csv", row.names = FALSE)
getwd()



# Day 3 Task 1

#--- Run summary statistics for age, weight, and height
```

```
summary(AGEP_A, WEIGHTLBTC_A, HEIGHTTC_A)

#--- Install and load mode

install.packages("modeest")
library(modeest)

#--- Run summary statistics, standard deviation, and mode for age, weight and height

summary(AGEP_A)
sd(AGEP_A)
mfv(AGEP_A)

summary(WEIGHTLBTC_A)
sd(WEIGHTLBTC_A)
mfv(WEIGHTLBTC_A)

summary(HEIGHTTC_A)
sd(HEIGHTTC_A)
mfv(HEIGHTTC_A)


#--- Create base R histogram for age, weight, and height


hist(AGEP_A)

hist(WEIGHTLBTC_A)

hist(HEIGHTTC_A)


#--- Create ggplot2 histogram for age, weight, and height


ggplot(nhis_clean_1_, aes(x=AGEP_A)) +
  geom_histogram(binwidth=2, fill="steelblue", color="black") +
  ggtitle("Histogram of Age") +
  xlab("Age") +
  ylab("Count")

ggplot(nhis_clean_1_, aes(x=WEIGHTLBTC_A)) +
  geom_histogram(binwidth=2, fill="steelblue", color="black") +
  ggtitle("Histogram of WEIGHT") +
  xlab("WEIGHT") +
  ylab("Count")

ggplot(nhis_clean_1_, aes(x=HEIGHTTC_A)) +
  geom_histogram(binwidth=2, fill="steelblue", color="black") +
  ggtitle("Histogram of HEIGHT") +
  xlab("HEIGHT") +
  ylab("Count")


#---Create base R boxplot for age, weight, and height


boxplot(AGEP_A)

boxplot(WEIGHTLBTC_A)

boxplot(HEIGHTTC_A)
```

```
#--- Create ggplot2 boxplot for age, weight, and height


ggplot(nhis_clean_1_, aes(y = AGEP_A)) +
  geom_boxplot(fill = "pink") +
  scale_x_discrete() +
  labs(title = "Boxplot of Age",
       y = "HEIGHT Count")

ggplot(nhis_clean_1_, aes(y = WEIGHTLBTC_A)) +
  geom_boxplot(fill = "pink") +
  scale_x_discrete() +
  labs(title = "Boxplot of WEIGHT",
       y = "HEIGHT Count")

ggplot(nhis_clean_1_, aes(y = HEIGHTTC_A)) +
  geom_boxplot(fill = "pink") +
  scale_x_discrete() +
  labs(title = "Boxplot of HEIGHT",
       y = "HEIGHT Count")


#--- Create frequency tables for gender, race/ethnicity, education, general health status,
and quality of life

table(nhis_clean_1_$SEX_A)

table(nhis_clean_1_$HISPALLP_A)

table(nhis_clean_1_$EDUCP_A)

table(nhis_clean_1_$PHSTAT_A)

table(nhis_clean_1_$LSATIS4R_A)

vars <- c("SEX_A", "HISPALLP_A", "EDUCP_A", "PHSTAT_A", "LSATIS4R_A")

lapply(nhis_clean_1_[vars], table)


#--- Create bar plot for gender, race/ethnicity, education, health status, and  quality of
life


counts <- table(nhis_clean_1_$SEX_A)

barplot(counts,
        main = "Distribution of SEX_A",
        xlab = "Gender",
        ylab = "Frequency",
        col = "skyblue",
        las = 2)

counts <- table(nhis_clean_1_$HISPALLP_A)

barplot(counts,
        main = "Distribution of Race/Ethnicity",
        xlab = "Race/Ethnicity",
        ylab = "Frequency",
        col = "skyblue",
        las = 2)

counts <- table(nhis_clean_1_$EDUCP_A)
```

```r
barplot(counts,
        main = "Distribution of Education",
        xlab = "Education Level",
        ylab = "Frequency",
        col = "skyblue",
        las = 2)

counts <- table(nhis_clean_1_$PHSTAT_A)

barplot(counts,
        main = "Distribution of General Health Status",
        xlab = "Health Status",
        ylab = "Frequency",
        col = "skyblue",
        las = 2)

counts <- table(nhis_clean_1_$LSATIS4R_A)

barplot(counts,
        main = "Distribution of Quality of Life",
        xlab = "Quality of Life",
        ylab = "Frequency",
        col = "skyblue",
        las = 2)

#--- Create ggplot2 barplot for gender, race/ethnicity, education, health status, and
quality of life


ggplot(nhis_clean_1_, aes(x=factor(SEX_A))) +
  geom_bar(fill="tomato", color="black") +
  ggtitle("Bar Plot of Gender") +
  xlab("Gender") +
  ylab("frequency")

ggplot(nhis_clean_1_, aes(x=factor(HISPALLP_A))) +
  geom_bar(fill="tomato", color="black") +
  ggtitle("Bar Plot of Race/Ethnicity") +
  xlab("Race/Ethnicity") +
  ylab("frequency")

ggplot(nhis_clean_1_, aes(x=factor(EDUCP_A))) +
  geom_bar(fill="tomato", color="black") +
  ggtitle("Bar Plot of Education") +
  xlab("Education") +
  ylab("frequency")

ggplot(nhis_clean_1_, aes(x=factor(PHSTAT_A))) +
  geom_bar(fill="tomato", color="black") +
  ggtitle("Bar Plot of General Health Status") +
  xlab("General Health Status") +
  ylab("frequency")

ggplot(nhis_clean_1_, aes(x=factor(LSATIS4R_A))) +
  geom_bar(fill="tomato", color="black") +
  ggtitle("Bar Plot of Quality of Life") +
  xlab("Quality of Life") +
  ylab("frequency")



#Day 3 Task 2
```

```
#Quantitative vs. Qualitative: Create side-by-side boxplots (using both plot() and
ggplot2)
#to visualize the distribution of AGEP_A across different levels of SEX_A and EDUCP_A.
NHISclean <- read.csv(file.choose(), header = T)
head(NHISclean) #Display the
attach(NHISclean) #Attach NHISclean data for easier coding

library(ggplot2) #Load ggplot2 package to use plots
par(mfcol=c(1,2)) #Displays one row and two columns of graphs


#Recode values of Sex variable from integers to meaningful words
NHISclean$SEX_A_new[NHISclean$SEX_A==1]<- "Male"
NHISclean$SEX_A_new[NHISclean$SEX_A==2]<- "Female"

#Create Base R Boxplot of distribution of Age by Sex
boxplot(AGEP_A~SEX_A_new, data= NHISclean ,xlab = "Sex", ylab = "Age",
        main= "Distribution of Age by Sex", col= "lavender",
        names= c("Male","Female"))

#Create Base R Boxplot of distribution of Age by Education Level
boxplot(AGEP_A~EDUCP_A_recoded, xlab = "Education Level", ylab = "Age",
        main= "Distribution of Age by Education Level", col="indianred",
        names= c("less than HS","HS Grad",
                      "Some College","College Grad+"))

#Create ggplot2 boxplot of Age by Sex
ggplot(NHISclean, aes(x=SEX_A_new, y= AGEP_A))+
  geom_boxplot(aes(fill=factor(SEX_A_new)))+
  labs(title = "Distribution of Age by Sex", x="Sex", y="Age")+
  theme_minimal()

#Create ggplots2 Boxplot of distribution of Age by Education Level
ggplot(NHISclean, aes(x=EDUCP_A_recoded, y=AGEP_A))+
  geom_boxplot(aes(fill = factor(EDUCP_A_recoded)))+
  labs(title = "Distribution of Age by Education Level",
       x= "Education Level",
       y="Age")+
  theme_minimal()

#Recode General Health Status variable from integers to meaningful words
NHISclean$PHSTAT_A_new[NHISclean$PHSTAT_A==1] <-"Excellent"
NHISclean$PHSTAT_A_new[NHISclean$PHSTAT_A==2] <-"Very good"
NHISclean$PHSTAT_A_new[NHISclean$PHSTAT_A==3] <-"Good"
NHISclean$PHSTAT_A_new[NHISclean$PHSTAT_A==4] <-"Fair"
NHISclean$PHSTAT_A_new[NHISclean$PHSTAT_A==5] <-"Poor"

#Recode General Health variable from integers to meaningful words
NHISclean$LSATIS4R_A_new[NHISclean$LSATIS4R_A==1]<- "Very Satisfied"
NHISclean$LSATIS4R_A_new[NHISclean$LSATIS4R_A==2]<- "Satisfied"
NHISclean$LSATIS4R_A_new[NHISclean$LSATIS4R_A==3]<- "Dissatisfied"
NHISclean$LSATIS4R_A_new[NHISclean$LSATIS4R_A==4]<- "Very Dissatisfied"

#Create a clustered bar chart (using ggplot2) to show the relationship between
#PHSTAT_A (General Health) and LSATIS4R_A (Life Satisfaction).

ggplot(data=NHISclean, aes(x=PHSTAT_A_new, y=LSATIS4R_A_new, fill = LSATIS4R_A_new))+
  geom_bar(stat = "identity", position= position_dodge())+
  labs(title = "Cluster Barplot of the relationship between
       General Health and Life Satisfaction",
       x="General Health",
       y="Life Satisfaction")

  #scale_fill_discrete(name="Life Satisfaction",
  #                    labels= c("Very Satisfied","Satisfied","Dissatisfied","Very
dissatisfied"))


#Create a scatter plot (using both plot() and ggplot2) of HEIGHTTC_A vs. WEIGHTLBTC_A.

par(mfcol=c(1,1)) #Displays one row and one column of graphs

#Create Base R scatter plot of HEIGHTTC_A vs. WEIGHTLBTC_A
plot(HEIGHTTC_A,WEIGHTLBTC_A,
     xlab = "Height",
     ylab = "Weight",
     pch = 19,
     col=4,
     smooth=FALSE,  #remove smooth estimate
     regLine=FALSE, #remove linear estimate
     main = "Distribution of Height vs Weight")

#Create ggplot2 scatter plot of HEIGHTTC_A vs. WEIGHTLBTC_A
ggplot(NHISclean, aes(x=HEIGHTTC_A, y=WEIGHTLBTC_A))+
  geom_point(color="royalblue")+
  geom_smooth(method="lm")+     #observe regression line
  ggtitle("Distribution of Height vs Weight")+
  xlab("Height")+
  ylab("Weight")

#Calculate and report the correlation coefficient.
cor(HEIGHTTC_A,WEIGHTLBTC_A)
#Results= 0.4890768


# Day 4 Task 1

# Load required libraries
library(tidyverse) # Data for wrangling and visualization
install.packages("psych") # Data for matrix correlation
library(psych)
library(ggplot2) # Data for plotting
nhis_clean <- read_csv("data/nhis_clean.csv")

# Create output folder
dir.create("plots", showWarnings = FALSE) # Prevents saving output in an already existing
folder

# Scatter plot Height vs Weight
ggplot(nhis_clean, aes(x = HEIGHTTC_A, y = WEIGHTLBTC_A, color = SEX_A)) +
  geom_point() +
  facet_wrap(~EDUCP_A_recoded) + # You can switch to PHSTAT_A instead
  labs(
    title = "Height vs Weight by Sex and Education",
    x = "Height",
    y = "Weight"
  )

# Save scatter plot
    ggsave("plots/day4_scatter.png")

# Correlation scatter plot matrix
pairs.panels(
  nhis_clean %>% select(AGEP_A, WEIGHTLBTC_A, HEIGHTTC_A)
)
```