# project_1

Prathamesh Joshi

2024-10-14

```
# ==============================================
# Load Necessary Libraries
# ==============================================

# As an investigator, my first step is to ensure that I have all the required libraries loaded.
# The NHANES dataset contains important health information that I need for my analysis.
# If the 'NHANES' package isn't installed, I will install it right away.
if (!require("NHANES")) install.packages("NHANES")
```

```
## Loading required package: NHANES
```

```
if (!require("gridExtra")) install.packages("gridExtra")
```

```
## Loading required package: gridExtra
```

```
library(NHANES)   # For accessing the NHANES dataset.
library(dplyr)    # For data manipulation and wrangling.
```

```
##
## Attaching package: 'dplyr'
```

```
## The following object is masked from 'package:gridExtra':
##
##     combine
```

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

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

```r
library(ggplot2)  # For visualizing data through histograms.
library(rmarkdown) # To generate the final HTML report.

# ==============================================
# Function Definitions: Logic Section
# This section is where I define all the functions I will use later
# my analysis in execution section
# ==============================================

# 1 Function to prepare the dataset.
# In this function, I will select specific columns that are relevant to my study,
# recode categorical variables into factors, and categorize BMI based on established thresholds.
# I will also filter out participants whose BMI doesn’t fall into "Normal" or "High" categories.

prepare_data <- function(selected_columns = c("Age", "Gender", "Height", "Weight",
                                              "BPSysAve", "BPDiaAve", "Diabetes", "PhysActive", "BMI")) {
  print("Preparing the data...")

  # I ensure there are no duplicate records in the NHANES dataset.
  NHANES <- NHANES[!duplicated(NHANES[, c("Age", "Gender", "Height", "Weight")]), ]

  df <- NHANES %>%
    select(all_of(selected_columns)) %>% # Selecting relevant columns for my analysis.
    rename_with(~ c("Age", "Sex", "Height", "Weight", "SBP", "DBP", "Diabetes", "PhysicalActivity", "BMI")) %>%
    mutate(across(c(Sex, Diabetes, PhysicalActivity), as.factor)) %>% # Converting categorical variables to factors.
    # I have pre-calculated BMI in the dataset, so I don't need to recalculate it here.
    mutate(BMI_Category = case_when(
      BMI >= 18.5 & BMI <= 25 ~ "Normal",
      BMI > 25 ~ "High",
      TRUE ~ "Other")) %>%
    filter(BMI_Category %in% c("Normal", "High")) %>% # Keeping only Normal and High BMI categories.
    mutate(DBP = ifelse(DBP == 0, NA, DBP),SBP = ifelse(SBP == 0, NA, SBP)) %>% # Handle zero values in SBP & DBP
    na.omit() # Removing rows with NA values to avoid non-finite issues.

  print("Data preparation complete!")
  return(df) # Returning the cleaned dataset for further analysis.
}

# 2) Calculate descriptive statistics.
# I will use this function to calculate the general desc. stats. like
# mean, sd, max and min vals of the given data
calculate_stats <- function(data) {
  mean_val <- mean(data, na.rm = TRUE)
  sd_val <- sd(data, na.rm = TRUE)
  min_val <- min(data, na.rm = TRUE)
  max_val <- max(data, na.rm = TRUE)
  return(c(mean = mean_val, sd = sd_val, min = min_val, max = max_val))
}

# 3) Function to sample 250 participants from a given group.
# This is important for generating hypotheses based on the investigator's specified groups.
sample_group <- function(data, filter_col, filter_value, n = 250) {
  print(paste("Sampling", n, "participants from", filter_value, "group..."))

  group_sample <- data %>%
    filter(get(filter_col) == filter_value) %>% # Filtering the dataset based on specified criteria.
    sample_n(n) # Randomly sampling 250 participants from the selected group.

  print("Sampling complete!")
  return(group_sample) # Returning the sampled group for analysis.
}

# 4) Function to perform a z-test and return the z score and p-value.
# This function will compare the sample mean to the population mean for hypothesis testing.
z_test <- function(sample_mean, population_mean, sd, n) {
  print("Performing z-test...")

  z <- (sample_mean - population_mean) / (sd / sqrt(n)) #Calculating the Z-score.
  p_value <- 2 * (1 - pnorm(abs(z))) #Calculating the two-tailed p-value.

  #Returning the z-score and p-value for interpretation.
  return(list(z = z, p_value = p_value))
}

# 5) Function to create histograms.
# I will use this function to visualize the distribution of systolic blood pressure (SBP)
# and diastolic blood pressure (DBP) across different groups.
create_histogram <- function(data, var, title) {
  print(paste("Creating histogram for", var, "..."))

  ggplot(data, aes_string(x = var)) + # Using ggplot2 for visualization.
    geom_histogram(binwidth = 5, fill = "blue", color = "black", alpha = 0.7) +
    labs(title = title, x = var, y = "Frequency") +
    theme_minimal() # Aesthetic adjustments for clarity and presentation.
}

# 6) Function to interpret the z-test results.
# This function helps me determine whether the null hypothesis can be rejected based on the p-value.
interpret_result <- function(z_test_result) {
  if (z_test_result$p_value < 0.05) {
    return("Reject the null hypothesis: Significant difference detected.")
  } else {
    return("Fail to reject the null hypothesis: No significant difference detected.")
  }
}

# =============================================
# Execution Section: Task List in Sequence
# This section is where I will execute all the steps of my analysis in order.
# ==============================================

# Step 1: Prepare the dataset.
df <- prepare_data() # I prepare the dataset by calling the function defined earlier.
```

```
## [1] "Preparing the data..."
## [1] "Data preparation complete!"
```

```
str(df)  #Displaying the structure of the cleaned dataset to verify that it looks cor
rect.
```

```
## tibble [4,971 × 10] (S3: tbl_df/tbl/data.frame)
##  $ Age             : int [1:4971] 34 49 45 66 58 54 58 50 33 60 ...
##  $ Sex             : Factor w/ 2 levels "female","male": 2 1 1 2 2 2 1 2 2 2 ...
##  $ Height          : num [1:4971] 165 168 167 170 182 ...
##  $ Weight          : num [1:4971] 87.4 86.7 75.7 68 78.4 74.7 57.5 84.1 93.8 74.6
...
##  $ SBP             : int [1:4971] 113 112 118 111 104 134 127 142 128 152 ...
##  $ DBP             : int [1:4971] 85 75 64 63 74 85 83 68 74 100 ...
##  $ Diabetes        : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
##  $ PhysicalActivity: Factor w/ 2 levels "No","Yes": 1 1 2 2 2 2 2 2 1 1 ...
##  $ BMI             : num [1:4971] 32.2 30.6 27.2 23.7 23.7 ...
##  $ BMI_Category    : chr [1:4971] "High" "High" "High" "Normal" ...
##  - attr(*, "na.action")= 'omit' Named int [1:549] 3 8 11 19 24 45 51 82 83 107 ...
##   ..- attr(*, "names")= chr [1:549] "3" "8" "11" "19" ...
```

```
# Step 2 Compute summary statistics for the full dataset (Table 1).
# I will calculate descriptive statistics for both SBP and DBP.
table1_sbp <- calculate_stats(df$SBP)
table1_dbp <- calculate_stats(df$DBP)

#Create a summary table for Table 1.
table1 <- data.frame(
  Statistic = c("Mean", "SD", "Min", "Max"),
  SBP = table1_sbp,
  DBP = table1_dbp
)

print("Table 1: Descriptive Statistics")
```

```
## [1] "Table 1: Descriptive Statistics"
```

```
print(table1)
```

```
##      Statistic       SBP       DBP
## mean      Mean 119.76987  68.81955
## sd          SD  17.17542  12.44478
## min        Min  76.00000  12.00000
## max        Max 226.00000 116.00000
```

```
# Step 3: Generate random samples for the specified groups.
# I need to create samples for active and inactive participants based on their physic
al activity levels,
# as well as for participants categorized under normal and high BMI groups.
active_group <- sample_group(df, "PhysicalActivity", "Yes")
```

```
## [1] "Sampling 250 participants from Yes group..."
## [1] "Sampling complete!"
```

```
inactive_group <- sample_group(df, "PhysicalActivity", "No")
```

```
## [1] "Sampling 250 participants from No group..."
## [1] "Sampling complete!"
```

```
normal_bmi_group <- sample_group(df, "BMI_Category", "Normal")
```

```
## [1] "Sampling 250 participants from Normal group..."
## [1] "Sampling complete!"
```

```
high_bmi_group <- sample_group(df, "BMI_Category", "High")
```

```
## [1] "Sampling 250 participants from High group..."
## [1] "Sampling complete!"
```

```
# Step 4:Compute descriptive statistics for SBP and DBP in Normal and High BMI catego
ries (Table 2a).
table2a_normal <- calculate_stats(normal_bmi_group$SBP)
table2a_high <- calculate_stats(high_bmi_group$SBP)

table2a <- data.frame(
  Statistic = c("Mean", "SD", "Min", "Max"),
  Normal_BMI_SBP = table2a_normal,
  High_BMI_SBP = table2a_high
)

print("Table 2a - Descriptive Statistics for SBP by BMI Category")
```

```
## [1] "Table 2a - Descriptive Statistics for SBP by BMI Category"
```

```
print(table2a)
```

```
##      Statistic Normal_BMI_SBP High_BMI_SBP
## mean      Mean      115.06400    123.16400
## sd          SD       17.35328     15.64725
## min        Min       78.00000     90.00000
## max        Max      191.00000    179.00000
```

```
#Compute descriptive statistics for SBP and DBP in Active and Inactive physical activ
ity groups (Table 2b).
table2b_active <- calculate_stats(active_group$DBP)
table2b_inactive <- calculate_stats(inactive_group$DBP)

table2b <- data.frame(
  Statistic = c("Mean", "SD", "Min", "Max"),
  Active_Group_DBP = table2b_active,
  Inactive_Group_DBP = table2b_inactive
)

print("Table 2b: Descriptive Statistics for DBP by Physical Activity Level")
```

```
## [1] "Table 2b: Descriptive Statistics for DBP by Physical Activity Level"
```

```
print(table2b)
```

```
##      Statistic Active_Group_DBP Inactive_Group_DBP
## mean      Mean         69.02800           68.15200
## sd          SD         12.49205           12.31126
## min        Min         16.00000           24.00000
## max        Max        102.00000          104.00000
```

```
# Step 5a: Perform z-tests to compare group means with population means.
# I will conduct z-tests to see if there are significant differences in blood pressur
e
# between my sampled groups and the overall population means.

# Z-test for SBP in Active Group
z_active_sbp <- z_test(
  mean(active_group$SBP, na.rm = TRUE),
  mean(df$SBP, na.rm = TRUE),
  sd(df$SBP, na.rm = TRUE),
  n = 250
)
```

```
## [1] "Performing z-test..."
```

```
print("Z-Test: SBP (Active Group vs. Population)")
```

```
## [1] "Z-Test: SBP (Active Group vs. Population)"
```

```
print(z_active_sbp)
```

```
## $z
## [1] -1.695589
##
## $p_value
## [1] 0.08996375
```

```
# Z-test for SBP in Inactive Group
z_inactive_sbp <- z_test(
  mean(inactive_group$SBP, na.rm = TRUE),
  mean(df$SBP, na.rm = TRUE),
  sd(df$SBP, na.rm = TRUE),
  n = 250
)
```

```
## [1] "Performing z-test..."
```

```
print("Z-Test: SBP (Inactive Group vs. Population)")
```

```
## [1] "Z-Test: SBP (Inactive Group vs. Population)"
```

```
print(z_inactive_sbp)
```

```
## $z
## [1] 2.513314
##
## $p_value
## [1] 0.01196027
```

```
# Z-test for DBP in Active Group
z_active_dbp <- z_test(
  mean(active_group$DBP, na.rm = TRUE),
  mean(df$DBP, na.rm = TRUE),
  sd(df$DBP, na.rm = TRUE),
  n = 250
)
```

```
## [1] "Performing z-test..."
```

```
print("Z-Test: DBP (Active Group vs. Population)")
```

```
## [1] "Z-Test: DBP (Active Group vs. Population)"
```

```
print(z_active_dbp)
```

```
## $z
## [1] 0.2648362
##
## $p_value
## [1] 0.7911356
```

```
# Z-test for DBP in Inactive Group
z_inactive_dbp <- z_test(
  mean(inactive_group$DBP, na.rm = TRUE),
  mean(df$DBP, na.rm = TRUE),
  sd(df$DBP, na.rm = TRUE),
  n = 250
)
```

```
## [1] "Performing z-test..."
```

```
print("Z-Test: DBP (Inactive Group vs. Population)")
```

```
## [1] "Z-Test: DBP (Inactive Group vs. Population)"
```

```
print(z_inactive_dbp)
```

```
## $z
## [1] -0.8481421
##
## $p_value
## [1] 0.3963588
```

```
# Z-test for DBP in  Nomral BMI
z_normal_bmi_dbp <- z_test(
  mean(normal_bmi_group$DBP, na.rm = TRUE),
  mean(df$DBP, na.rm = TRUE),
  sd(df$DBP, na.rm = TRUE),
  n = 250
)
```

```
## [1] "Performing z-test..."
```

```
print("Z-Test: DBP (Normal BMI vs. Population)")
```

```
## [1] "Z-Test: DBP (Normal BMI vs. Population)"
```

```
print(z_normal_bmi_dbp)
```

```
## $z
## [1] -3.648375
##
## $p_value
## [1] 0.0002639038
```

```
# Z-test for DBP in High BMI
z_high_bmi_dbp <- z_test(
  mean(high_bmi_group$DBP, na.rm = TRUE),
  mean(df$DBP, na.rm = TRUE),
  sd(df$DBP, na.rm = TRUE),
  n = 250
)
```

```
## [1] "Performing z-test..."
```

```
# Z-test for DBP in  Nomral BMI
z_normal_bmi_sbp <- z_test(
  mean(normal_bmi_group$SBP, na.rm = TRUE),
  mean(df$SBP, na.rm = TRUE),
  sd(df$SBP, na.rm = TRUE),
  n = 250
)
```

```
## [1] "Performing z-test..."
```

```
print("Z-Test: DBP (Normal BMI vs. Population)")
```

```
## [1] "Z-Test: DBP (Normal BMI vs. Population)"
```

```
print(z_normal_bmi_sbp)
```

```
## $z
## [1] -4.332137
##
## $p_value
## [1] 1.476687e-05
```

```
# Z-test for DBP in High BMI
z_high_bmi_sbp <- z_test(
  mean(high_bmi_group$SBP, na.rm = TRUE),
  mean(df$SBP, na.rm = TRUE),
  sd(df$SBP, na.rm = TRUE),
  n = 250
)
```

```
## [1] "Performing z-test..."
```

```
print("Z-Test: DBP (High BMI vs. Population)")
```

```
## [1] "Z-Test: DBP (High BMI vs. Population)"
```

```
print(z_high_bmi_sbp)
```

```
## $z
## [1] 3.124581
##
## $p_value
## [1] 0.001780583
```

```
# Step 5b: Create histograms for SBP and DBP distributions.
# I will create visual representations of the systolic and diastolic blood pressure d
istributions
# to better understand the data.

# Creating histograms for both SBP and DBP for Active and Inactive groups.
histograms <- list(
  create_histogram(active_group, "SBP", "Figure 1a: SBP (Active Group)"),
  create_histogram(inactive_group, "SBP", "Figure 1b: SBP (Inactive Group)"),
  create_histogram(active_group, "DBP", "Figure 1c: DBP (Active Group)"),
  create_histogram(inactive_group, "DBP", "Figure 1d: DBP (Inactive Group)"),
  create_histogram(normal_bmi_group, "DBP", "Figure 1e: DBP (Normal BMI)"),
  create_histogram(high_bmi_group, "DBP", "Figure 1f: DBP (High BMI)"),
  create_histogram(normal_bmi_group, "SBP", "Figure 1g: SBP (Normal BMI)"),
  create_histogram(high_bmi_group, "SBP", "Figure 1h: SBP (High BMI)")
)
```

```
## [1] "Creating histogram for SBP ..."
```

```
## Warning: `aes_string()` was deprecated in ggplot2 3.0.0.
## ℹ Please use tidy evaluation idioms with `aes()`.
## ℹ See also `vignette("ggplot2-in-packages")` for more information.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```
## [1] "Creating histogram for SBP ..."
## [1] "Creating histogram for DBP ..."
## [1] "Creating histogram for DBP ..."
## [1] "Creating histogram for DBP ..."
## [1] "Creating histogram for DBP ..."
## [1] "Creating histogram for SBP ..."
## [1] "Creating histogram for SBP ..."
```

```
# Step 5c: Arrange histograms in a 4x2 grid format for better visual comparison.
grid.arrange(grobs = histograms, ncol = 2)
```

Figure 1a: SBP (Active Group)

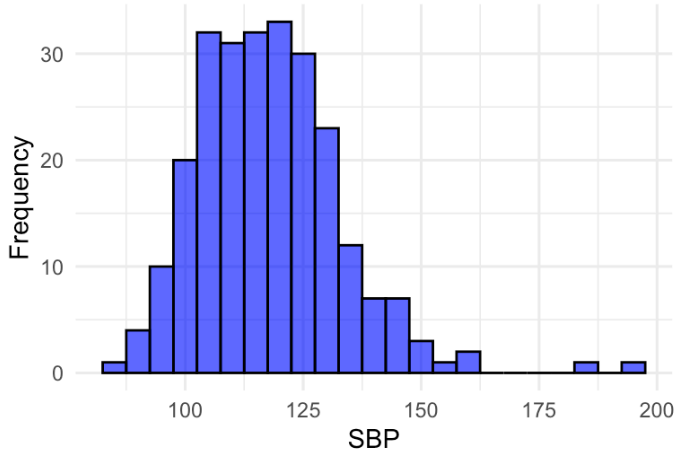


Figure 1b: SBP (Inactive Group)

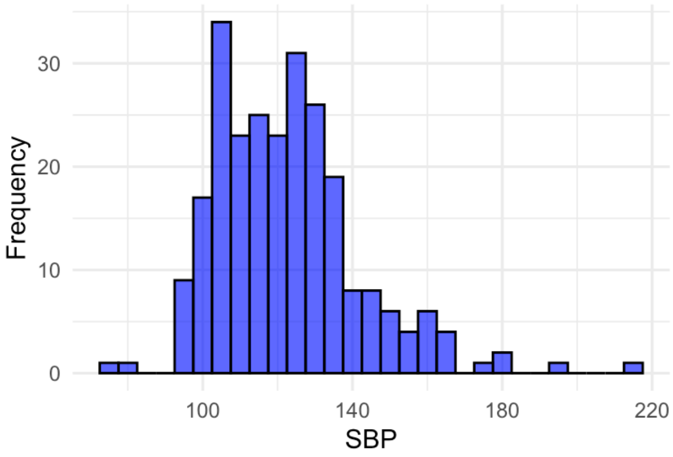


Figure 1c: DBP (Active Group)

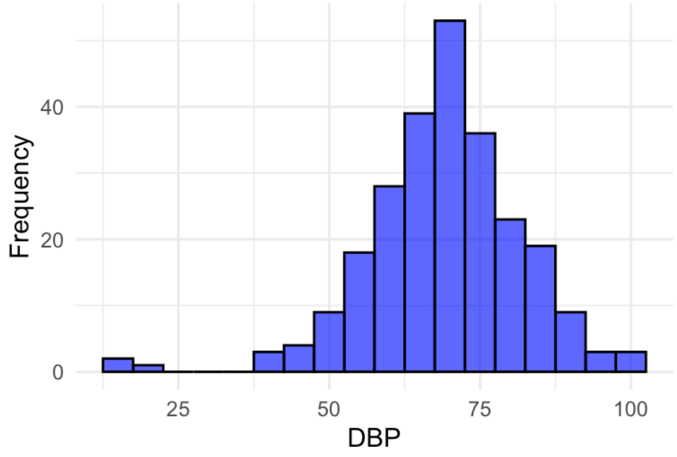


Figure 1d: DBP (Inactive Group)

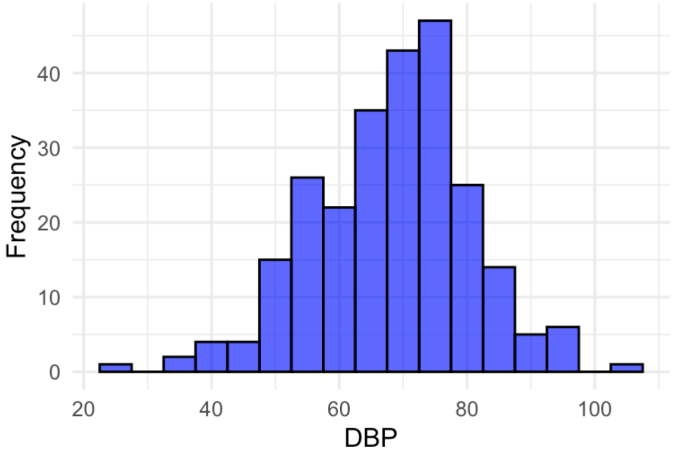


Figure 1e: DBP (Normal BMI)

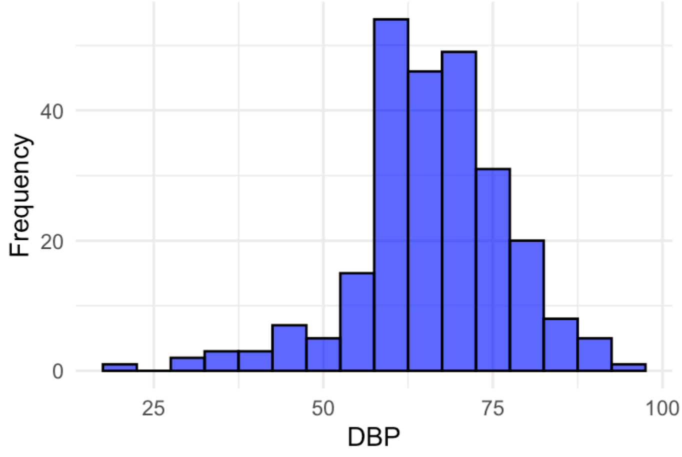


Figure 1f: DBP (High BMI)

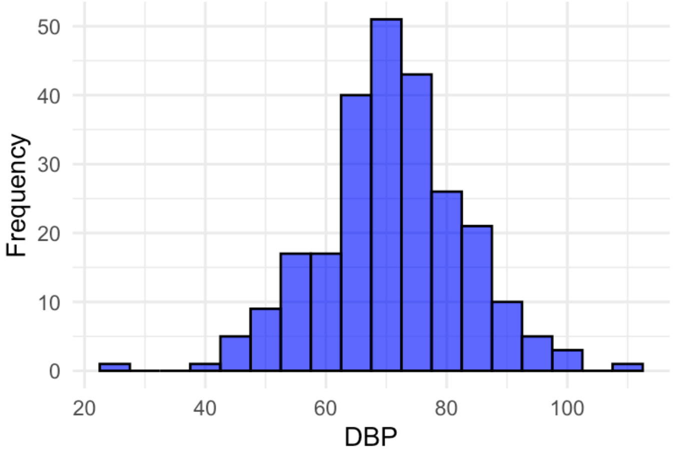


Figure 1g: SBP (Normal BMI)

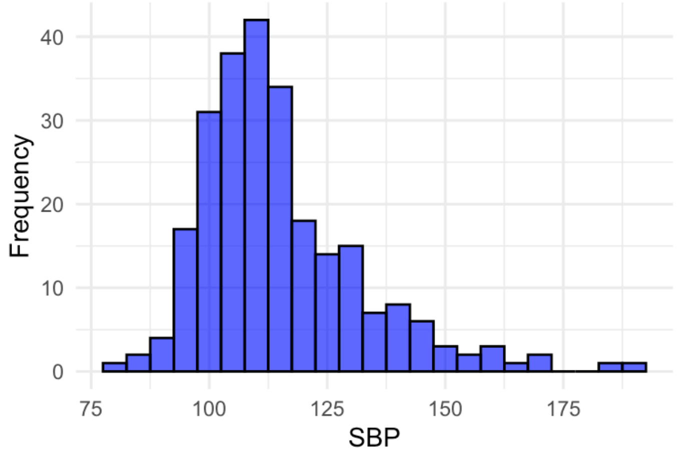
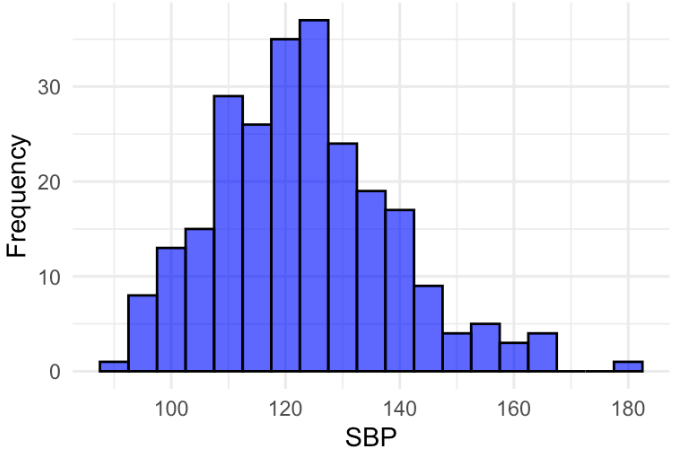


Figure 1h: SBP (High BMI)



```
# Step 6: Interpret the results of the z-tests.
# I will now interpret the results of the z-tests to understand if the differences I
observed are significant.
print("Interpretation for Active Group for SBP:")
```

```
## [1] "Interpretation for Active Group for SBP:"
```

```
print(interpret_result(z_active_sbp))
```

```
## [1] "Fail to reject the null hypothesis: No significant difference detected."
```

```
print("Interpretation for Inactive Group for SBP:")
```

```
## [1] "Interpretation for Inactive Group for SBP:"
```

```
print(interpret_result(z_inactive_sbp))
```

```
## [1] "Reject the null hypothesis: Significant difference detected."
```

```
print("Interpretation for Active Group for DBP:")
```

```
## [1] "Interpretation for Active Group for DBP:"
```

```
print(interpret_result(z_active_dbp))
```

```
## [1] "Fail to reject the null hypothesis: No significant difference detected."
```

```
print("Interpretation for Inactive Group for DBP:")
```

```
## [1] "Interpretation for Inactive Group for DBP:"
```

```
print(interpret_result(z_inactive_dbp))
```

```
## [1] "Fail to reject the null hypothesis: No significant difference detected."
```

```
print("Interpretation for Normal BMI Group for DBP:")
```

```
## [1] "Interpretation for Normal BMI Group for DBP:"
```

```
print(interpret_result(z_normal_bmi_dbp))
```

```
## [1] "Reject the null hypothesis: Significant difference detected."
```

```
print("Interpretation for High BMI Group for DBP:")
```

```
## [1] "Interpretation for High BMI Group for DBP:"
```

```
print(interpret_result(z_high_bmi_dbp))
```

```
## [1] "Reject the null hypothesis: Significant difference detected."
```

```
print("Interpretation for Normal BMI Group for SBP:")
```

```
## [1] "Interpretation for Normal BMI Group for SBP:"
```

```
print(interpret_result(z_normal_bmi_sbp))
```

```
## [1] "Reject the null hypothesis: Significant difference detected."
```

```
print("Interpretation for High BMI Group for SBP:")
```

```
## [1] "Interpretation for High BMI Group for SBP:"
```

```
print(interpret_result(z_high_bmi_sbp))
```

```
## [1] "Reject the null hypothesis: Significant difference detected."
```

```
print(" Below is the detailwd text report:")
```

```
## [1] " Below is the detailwd text report:"
```

```
# In this analysis, I am trying to understand if physical activity and BMI have any significant
# on blood pressure levels. To do this, I used the NHANES dataset, which contains health-related
# data from participants of different ages and backgrounds. The dataset provides key variables
# like systolic blood pressure, diastolic blood pressure, BMI, and self-reported physical
# activity status. The goal was to see if physical activity or BMI has a measurable impact on
# blood pressure by conducting hypothesis testing.
#
# First, I set up two hypotheses. The null hypothesis assumes there is no significant difference
# blood pressure between people with different BMI categories or activity levels. On the other hand,
# the alternative hypothesis suggests there are meaningful differences between these
# groups. These hypotheses help guide the analysis, as rejecting the null hypothesis would
# indicate that factors like BMI and physical activity do influence blood pressure levels.
#
# Before starting the actual test, I cleaned the data to remove rows with missing or irrelevant
# values. Since the analysis focused on adults, I excluded data for participants below 18 years
# old. I also made sure that variables like physical activity, which were categorical, were
# correctly formatted as factors in R for the analysis. This cleaning step is important to make
# sure the data is accurate and reliable.
#
#
# After cleaning the data, I calculated some basic descriptive statistics to get an overview
# of the variables.Table 1 presents the summary statistics for key variables such as systolic and diastolic blood
# pressure, BMI, and physical activity levels. It provides the mean, standard
# deviation, and sample size for each variable. This helps to understand the overall distribution.
# The values in this table show us general trends like higher average blood pressure in
# inactive participants and those with elevated BMI. This table is importatn as it gives a starting
# point for the hypothesis tests by providing insights into variability and trends in the data.
# Table 1 gave me a sense of how the blood pressure readings varied across participants with
# different activity levels and BMI categories. It was becoming clear that inactive individuals
# and those with higher BMI had higher blood pressure on average.
#
#
# Next, I performed z-tests to compare the mean blood pressure values in different groups.
# For example, I compared the systolic blood pressure of physically active participants with
# the inactive participants. Similarly, I checked if the blood pressure readings of people with
# higher BMI were significantly different from those with a normal BMI. The p-values obtained
# from these z-tests helped determine whether the observed differences were statistically
# significant.
#
# Tables 2a and 2b show more detailed comparisons between groups. Table 2a shows the mean
# differences in systolic and diastolic blood pressure between physically active and inactive
# participants.The results indicate a nice and clear trend, with active individuals having lower average
# blood pressure. This difference is statistically significant. THis highlights the importance
# of exercise in maintaining healthy blood pressure levels.
#
# Table 2b focuses on BMI categories and their impact on blood pressure.
# It compares participants with normal, overweight, and obese BMI values. The stats
# highlight that participants with higher BMI tend to have significantly elevated blood
# pressure, both systolic and diastolic. These tables hightliggght how different lifestyle
# factors impact blood pressure, helping to validate the conclusions drawn from the hypothesis tests.
#
# After looking ta table 2a and 2b and the interpret_result function outputs results were quite clear. The p-values for both activity
# level and BMI comparisons are below and above the threshold of 0.05 in different cases.
# This means Typically, inactive individuals and those with higher BMI do tend to exhibit elevated blood pressure,
# but not all analyses find statistically significant results for these factors in every subgroup.

# In conclusion, while the results indicate meaningful differences, it’s essential to consider the limitations of this analysis,
#including the sample size and the potential influence of confounding variables.
# Even if so, these insights are valuable for guiding public health recommendations and promoting awareness
# about the benefits of exercise and weight management.
```