Varun Sudhir 21BDS0040

Exploratory Data Analysis Lab Exp 7

Use statistical techniques to identify outlier data for the given data set.

For this experiment,we will utilize the NHANES dataset. The NHANES (National Health and Nutrition Examination Survey) dataset is a large, real-world dataset that comes from a program of studies conducted by the National Center for Health Statistics (NCHS). The NHANES dataset is often used in health and epidemiological studies and contains various missing values across its columns.

We will be considering the BMI column of the dataset and find the outliers in this particular column

```
> print(BMI)
  [1] 32.22 32.22 32.22 15.30 30.57 16.82 20.64 27.24 27.24 27.24 23.67 23.69 26.03 19.20 26.22 26.60 27.40 28.54 25.84
 [20] 24.74 19.73 19.73 20.66 36.32 36.32 35.84 24.32 25.95 31.43 31.43 27.18 21.00 25.79 25.79 29.13 30.60 30.60 23.34
 [39] 22.85 22.85 26.46 26.46 26.46 26.46 25.45 21.16 46.69 20.15 27.06 37.33 37.33 15.59 15.59 25.54 24.98 22.63 14.35
 [58] 37.92 37.92 37.92    NA 18.16 25.52 28.96 28.96 32.49 32.49 32.49 18.35 16.24 16.24 28.48 28.48 19.41 36.28 25.87
 [77] 25.87 25.87 28.60 21.03 21.03 21.03 30.90 30.90 30.90 30.90 31.51 31.51 27.74 27.25 27.25 24.53 29.83 22.81 29.27
```

```r
# Varun Sudhir 21BDS0040
# Select the 'BMI' column from the dataset
BMI <- NHANES$BMI

# Remove missing values
BMI_clean <- BMI[!is.na(BMI)]

# Plotting a box plot for BMI to identify outliers
ggplot(data.frame(BMI_clean), aes(x = "", y = BMI_clean)) +
  geom_boxplot(outlier.colour = "red", outlier.shape = 16, outlier.size = 2) +
  labs(title = "Box Plot of BMI with Outliers ( 21BDS0040 )", y = "BMI") +
  theme_minimal()

# Function to detect outliers using the IQR method
detect_outliers_IQR <- function(data) {
  Q1 <- quantile(data, 0.25)  # First quartile (25th percentile)
  Q3 <- quantile(data, 0.75)  # Third quartile (75th percentile)
  IQR_value <- Q3 - Q1        # Interquartile Range (IQR)

  # Define the outlier thresholds
  lower_bound <- Q1 - 1.5 * IQR_value
  upper_bound <- Q3 + 1.5 * IQR_value

  # Identify outliers
  outliers <- data[data < lower_bound | data > upper_bound]
```

```
  return(list("Outliers" = outliers, "Lower Bound" = lower_bound, "Upper
Bound" = upper_bound))
}

# Detect outliers in the BMI column
outliers_result <- detect_outliers_IQR(BMI_clean)

# Print the detected outliers
cat("Outliers in BMI: ", outliers_result$Outliers, "\n")
print("Varun Sudhir 21BDS0040")
```

Output:

```
> # Print the detected outliers
> cat("Outliers in BMI: ", outliers_result$Outliers, "\n")
Outliers in BMI:  46.69 48.91 51.33 48.22 48.22 48.22 45.82 45.27 81.25 81.25 52.08 45.77 48.12 51.54 51.54 52.03 45.75 52.65 52.86 48.2 48.2
47.11 45.65 45.65 45.65 45.65 49.47 49.47 46.26 45.96 45.91 63.89 47.06 65.19 58.18 45.72 45.72 53.78 52.34 48.28 46.62 49.37 48.22 48.22 48.
22 48.87 48.87 48.87 53.83 45.72 46.75 66.96 45.65 46.05 49.46 47.99 46.31 44.87 47.35 46.03 55.95 47.87 45.45 46.26 47.11 46.27 45.47 47.94
48.68 46.42 65.62 48.09 44.98 45.39 59.2 63.91 63.91 67.83 55.07 46.45 54.94 47.13 45.17 46.97 46.97 46.92 46.29 55.6 50.52 53.27 47.49 61.01
51.38 68.63 45.93 47.13 46.98 45.39 45.2 48.2 48.2 48.3 48.6 54.4 54.4 54.4 54.4 54.2 59.1 47.5 50 47.2 47.7 45.8 45.8 47.7 47.7 47.7 44.9 4
5.2 45.2 47 53.1 56.8 56.8 56.8 49.7 46.9 46.9 46.9 46.9 45.9 53.5 56.5 59.2 59.2 54.1 54.1 54.1 46.4 52.8 50.4 50.4 49.9 47.1 54.4 47.1 47.1
45.5 48.7 48.7 56.8 56.8 62.8 48.8 55 47.5 69 52.3 50.7 50.9 50.9 45.3 45.3 47.3 46.7 46.7 46.7 63.3 48 44.9 80.6 45.5 47.6 47.6 45.7 45.7 4
5.7 45.7 45.7 44.9 45.4 45.3 45.3 45.3 48
> print("Varun Sudhir 21BDS0040")
[1] "Varun Sudhir 21BDS0040"
>
```