Alcohol Consumption Data Analysis

Gopinath Achuthan

2024-09-22

## Description of the data

This dataset measures the average alcohol consumption per person in different countries, specifically focusing on beer, spirits, and wine servings. It provides insights into drinking habits globally and can help address research questions related to cultural consumption patterns, public health implications, and economic factors surrounding alcohol production and sales. The data is stored in a CSV format, which is a delimited text file where each line represents a row of data, and the values are separated by commas.

## Reading the data into R

# Load the necessary package  
suppressMessages(library(readr)) # for reading csv files into a dataframe  
suppressMessages(suppressWarnings(library(dplyr))) # for data cleaning  
suppressMessages(suppressWarnings(library(knitr))) # for displaying table  
  
# Read the CSV file and assign it to a dataframe  
url <- 'https://raw.githubusercontent.com/fivethirtyeight/data/refs/heads/master/alcohol-consumption/drinks.csv'  
alcohol\_data <- read\_csv(url,   
 col\_names = TRUE,  
 show\_col\_types = FALSE)  
  
# The col\_names = TRUE argument in read\_csv() indicates that the first row of the CSV file contains the column names  
# Setting show\_col\_types = FALSE in read\_csv() suppresses the display of column type information for a cleaner output  
  
# Display the raw data  
kable(head(alcohol\_data, 5),  
 caption = '<div style="text-align: center; color: turquoise;">Top 5 Raw Data</div>',  
 align = 'c'  
)

Table:

Top 5 Raw Data

| country | beer\_servings | spirit\_servings | wine\_servings | total\_litres\_of\_pure\_alcohol |
| --- | --- | --- | --- | --- |
| Afghanistan | 0 | 0 | 0 | 0.0 |
| Albania | 89 | 132 | 54 | 4.9 |
| Algeria | 25 | 0 | 14 | 0.7 |
| Andorra | 245 | 138 | 312 | 12.4 |
| Angola | 217 | 57 | 45 | 5.9 |

## Clean the data

# Clean the data: Renaming columns and filtering the column by Keeping only countries with positive alcohol consumption  
cleaned\_data <- alcohol\_data %>%  
 rename(  
 country\_name = country,  
 beer\_servings\_per\_person = beer\_servings,  
 spirit\_servings\_per\_person = spirit\_servings,  
 wine\_servings\_per\_person = wine\_servings,  
 total\_alcohol\_litres\_per\_person = total\_litres\_of\_pure\_alcohol  
 ) %>%  
 filter(total\_alcohol\_litres\_per\_person > 0)  
  
# Display the cleaned data  
kable(head(cleaned\_data, 5),  
 caption = '<div style="text-align: center; color: turquoise;">Top 5 Cleaned Data</div>',  
 align = 'c'  
)

Table:

Top 5 Cleaned Data

| country\_name | beer\_servings\_per\_person | spirit\_servings\_per\_person | wine\_servings\_per\_person | total\_alcohol\_litres\_per\_person |
| --- | --- | --- | --- | --- |
| Albania | 89 | 132 | 54 | 4.9 |
| Algeria | 25 | 0 | 14 | 0.7 |
| Andorra | 245 | 138 | 312 | 12.4 |
| Angola | 217 | 57 | 45 | 5.9 |
| Antigua & Barbuda | 102 | 128 | 45 | 4.9 |

## Characteristics of the data

# Create a characteristics table  
column\_info <- data.frame(  
 Column\_Number = 1:3,  
 Column\_Name = c("country\_name", "beer\_servings\_per\_person", "total\_alcohol\_litres\_per\_person"),  
 Description = c("Name of the country",   
 "Average beer serving sizes per person",   
 "Average total pure alcohol litres per person")  
)  
  
# column alignment: "l" for left, "c" for center, "r" for right  
alignment <- c("c", "l", "l") # Center for column 1, left for columns 2 and 3  
  
# Display the characteristics table using kable  
kable(column\_info,   
 caption = '<div style="text-align: center; color: turquoise;">Characteristics Table</div>',  
 align = alignment)

Table:

Characteristics Table

| Column\_Number | Column\_Name | Description |
| --- | --- | --- |
| 1 | country\_name | Name of the country |
| 2 | beer\_servings\_per\_person | Average beer serving sizes per person |
| 3 | total\_alcohol\_litres\_per\_person | Average total pure alcohol litres per person |

## Subset and Summary

# Create a subset of the dataframe with two selected columns  
subset\_data <- cleaned\_data[, c("beer\_servings\_per\_person", "total\_alcohol\_litres\_per\_person")]  
  
# Function to calculate mode  
get\_mode <- function(x) {  
 ux <- unique(x)  
 ux[which.max(tabulate(match(x, ux)))]  
}  
  
# Summarize the two columns  
summary\_stats <- data.frame(  
 Min = c(  
 min(subset\_data$beer\_servings\_per\_person, na.rm = TRUE),  
 min(subset\_data$total\_alcohol\_litres\_per\_person, na.rm = TRUE)  
 ),  
 Max = c(  
 max(subset\_data$beer\_servings\_per\_person, na.rm = TRUE),  
 max(subset\_data$total\_alcohol\_litres\_per\_person, na.rm = TRUE)  
 ),  
 Mean = c(  
 mean(subset\_data$beer\_servings\_per\_person, na.rm = TRUE),  
 mean(subset\_data$total\_alcohol\_litres\_per\_person, na.rm = TRUE)  
 ),  
 Mode = c(  
 get\_mode(subset\_data$beer\_servings\_per\_person),  
 get\_mode(subset\_data$total\_alcohol\_litres\_per\_person)  
 )  
)  
  
# Set row names for clarity  
rownames(summary\_stats) <- c("<b>Beer Servings</b>", "<b>Total Pure Alcohol</b>")  
  
# assign summary stats to a new object  
summary\_result <- summary\_stats  
  
# Display the summary results using kable  
kable(summary\_result,   
 caption = '<div style="text-align: center; color: turquoise;">Summary Statistics</div>',  
 align = 'c')

Table:

Summary Statistics

|  | Min | Max | Mean | Mode |
| --- | --- | --- | --- | --- |
| Beer Servings | 0.0 | 376.0 | 113.827778 | 5.0 |
| Total Pure Alcohol | 0.1 | 14.4 | 5.057778 | 0.1 |