Sewani-Sahil-Final Project

Sahil Sewani

2023-10-03

Description of data

The "Global Blood Type Distribution Dataset" compiles worldwide blood type information. It offers a broad view of ABO blood group prevalence across diverse populations and regions. This dataset is a valuable tool for researchers, healthcare professionals, and data analysts to study global blood type diversity. It can support a range of studies, including cross-cultural research and medical analysis. Researchers can use this raw dataset to uncover significant trends and correlations related to blood types in different countries. The data will be explored in a quantitative manner and displayed through bar-plots and histograms.

Link to dataset: https://www.kaggle.com/datasets/kamilenovaes/global-blood-type-distribution

GitHub Repository: https://github.com/Sahil-Sewani/bloodtypes

Finding the most common blood type

This code reads blood type data from a CSV file, processes it to exclude certain columns, calculates the total frequency of each blood type, and then identifies and prints the most common blood type.

```
# Load the tidyverse package
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.3     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become error
```

```
# Read the data with backticks around column names
blood_type_data <- read.csv("/Users/sahil_sewani/Documents/VCU/Fall 2023/Biostatistical Computing (BIOS

# Exclude the first two columns (Country and Population) for analysis and convert to a tibble
blood_type_data <- as_tibble(blood_type_data) %>%
  select(-Country, -Population)

# Calculate the total frequency for each blood type
blood_type_frequencies <- colSums(blood_type_data)
```

```r
### Find the most common blood type
most_common_blood_type <- names(blood_type_frequencies)[which.max(blood_type_frequencies)]

cat("The most common blood type worldwide is:", most_common_blood_type, "\n")
```

```
## The most common blood type worldwide is: O+
```

Calculate the average of each blood type

This code loads necessary R packages, installs and loads the 'bloodtypepackage', uses it to calculate the average frequency of each blood type from a CSV file, and then prints the result to the console.

```r
# Load the tidyverse package
library(tidyverse)

# Load the 'bloodtypepackage'
install.packages('bloodtypepackage')
```

```
## Warning: package 'bloodtypepackage' is not available for this version of R
##
## A version of this package for your version of R might be available elsewhere,
## see the ideas at
## https://cran.r-project.org/doc/manuals/r-patched/R-admin.html#Installing-packages
```

```r
library(bloodtypepackage)

averages_of_types <- bloodtypepackage::calculate_blood_type_averages("/Users/sahil_sewani/Documents/VCU/
```

```
## The most common blood type worldwide is: O+
```

```r
print(averages_of_types)
```

```
## # A tibble: 8 x 2
##   BloodType AverageFrequency
##   <chr>                <dbl>
## 1 O+                  40.4
## 2 A+                  29.7
## 3 B+                  16.4
## 4 AB+                  4.83
## 5 O-                   3.92
## 6 A-                   3.28
## 7 B-                   1.33
## 8 AB-                  0.496
```

Calculate the summary statistics of each blood type

this code loads necessary R packages, installs and loads the 'bloodtypepackage', uses it to calculate summary statistics for each blood type from a CSV file, and then prints the result to the console.

```r
# Load the tidyverse package
library(tidyverse)

# Load the 'bloodtypepackage'
install.packages('bloodtypepackage')
```

```
## Warning: package 'bloodtypepackage' is not available for this version of R
##
## A version of this package for your version of R might be available elsewhere,
## see the ideas at
## https://cran.r-project.org/doc/manuals/r-patched/R-admin.html#Installing-packages
```

```r
library(bloodtypepackage)

summary_statistics_types <- bloodtypepackage::calculate_blood_type_summary_stats("/Users/sahil_sewani/D
```

```
##        O+              A+              B+             AB+
##  Min.   :25.50   Min.   :14.00   Min.   : 4.72   Min.   : 0.500
##  1st Qu.:32.08   1st Qu.:25.82   1st Qu.:10.00   1st Qu.: 2.925
##  Median :38.17   Median :30.00   Median :15.00   Median : 4.295
##  Mean   :40.35   Mean   :29.67   Mean   :16.40   Mean   : 4.827
##  3rd Qu.:46.82   3rd Qu.:34.85   3rd Qu.:21.23   3rd Qu.: 6.300
##  Max.   :75.00   Max.   :46.30   Max.   :36.80   Max.   :14.700
##
##        O-              A-              B-             AB-
##  Min.   : 0.060   Min.   :0.040   Min.   :0.010   Min.   :0.0100
##  1st Qu.: 1.790   1st Qu.:1.000   1st Qu.:0.540   1st Qu.:0.1500
##  Median : 4.000   Median :2.700   Median :1.250   Median :0.4000
##  Mean   : 3.917   Mean   :3.277   Mean   :1.334   Mean   :0.4963
##  3rd Qu.: 6.000   3rd Qu.:6.000   3rd Qu.:2.000   3rd Qu.:0.9100
##  Max.   :13.000   Max.   :8.000   Max.   :3.130   Max.   :1.2000
##  NA's   :1        NA's   :1       NA's   :1       NA's   :1
```

Analysis of results: For O+ blood type, the frequencies range from a minimum of 25.50 to a maximum of 75.00. The data is spread across the quartiles as follows: the first quartile (25th percentile) lies at 32.08, the median (50th percentile) is 38.17, and the third quartile (75th percentile) is 46.82. The mean frequency, representing the average, is approximately 40.35.

A+ blood type shows frequencies ranging from a minimum of 14.00 to a maximum of 46.30. The first quartile is 25.82, the median is 30.00, and the third quartile is 34.85. The mean frequency is around 29.67.

B+ blood type exhibits frequencies with a minimum of 4.72 and a maximum of 36.80. The first quartile is 10.00, the median is 15.00, and the third quartile is 21.23. The mean frequency is approximately 16.40.

AB+ blood type has frequencies ranging from a minimum of 0.500 to a maximum of 14.700. The first quartile is 2.925, the median is 4.295, and the third quartile is 6.300. The mean frequency is about 4.827.

For O- blood type, frequencies range from a minimum of 0.060 to a maximum of 13.000. The first quartile is 1.790, the median is 4.000, and the third quartile is 6.000. The mean frequency is approximately 3.917.

A- blood type displays frequencies with a minimum of 0.040 and a maximum of 8.000. The first quartile is 1.000, the median is 2.700, and the third quartile is 6.000. The mean frequency is around 3.277.

B- blood type exhibits frequencies ranging from a minimum of 0.010 to a maximum of 3.130. The first quartile is 0.540, the median is 1.250, and the third quartile is 2.000. The mean frequency is approximately 1.334.

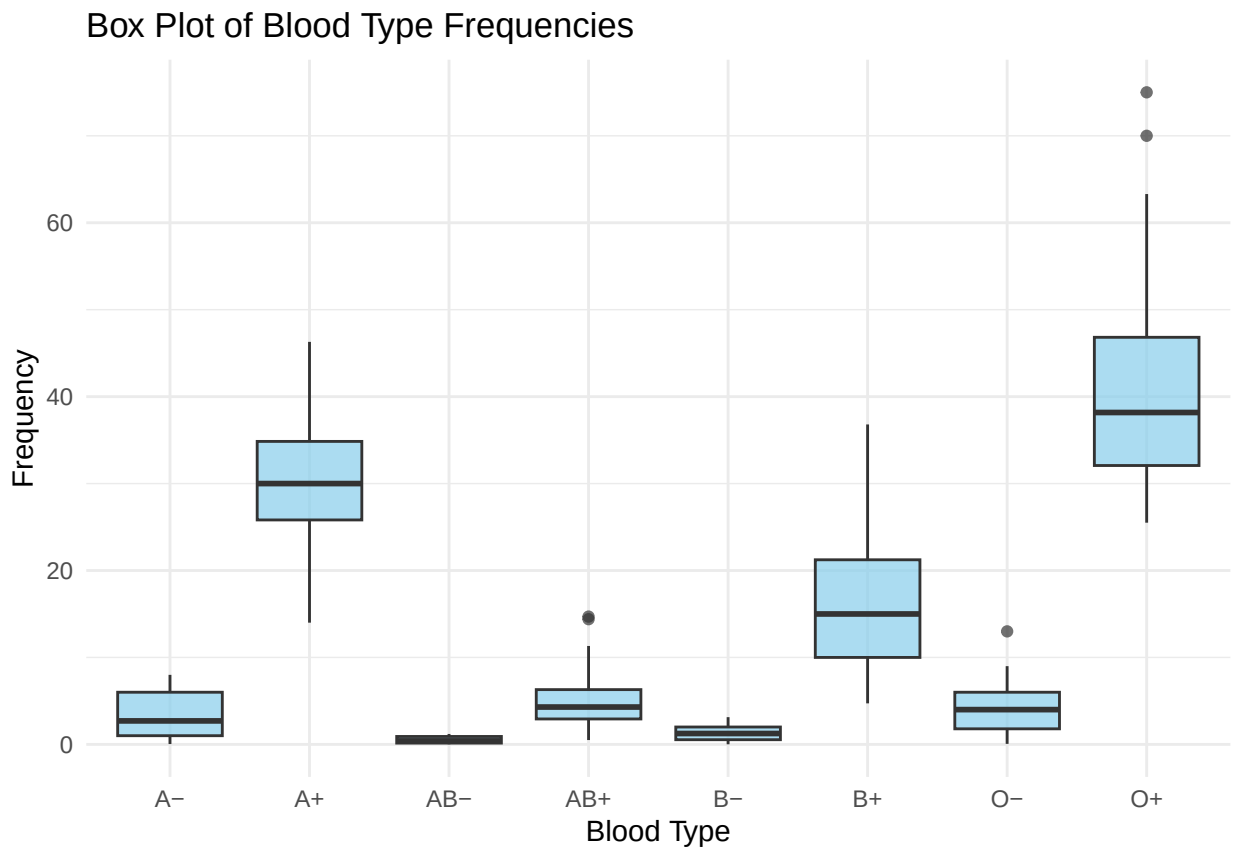
Finally, AB- blood type shows frequencies with a minimum of 0.010 and a maximum of 1.200. The first quartile is 0.150, the median is 0.400, and the third quartile is 0.910. The mean frequency is about 0.4963. Please note that there may be missing values in the dataset, as indicated by the presence of "NA's" in the summary.

Create box-plots for each blood type

This code defines a function named create_blood_type_boxplots which generates box plots for blood type frequencies.

```
# Load necessary libraries
library(tidyverse)

# Load the data from the CSV file
blood_type_data <- read.csv("/Users/sahil_sewani/Documents/VCU/Fall 2023/Biostatistical Computing (BIOS

# Manually set the column names
colnames(blood_type_data) <- c("Country", "Population", "O+", "A+", "B+", "AB+", "O-", "A-", "B-", "AB-

### Create box plots for each blood type (function)
create_blood_type_boxplots <- function(data) {
  data %>%
    pivot_longer(cols = -c(Country, Population), names_to = "Blood_Type", values_to = "Frequency") %>%
    ggplot(aes(x = Blood_Type, y = Frequency)) +
    geom_boxplot(fill = "skyblue", alpha = 0.7) +
    labs(title = "Box Plot of Blood Type Frequencies",
         x = "Blood Type",
         y = "Frequency") +
    theme_minimal()
}

### initialize box-plot function
create_blood_type_boxplots(blood_type_data)
```

```
## Warning: Removed 4 rows containing non-finite values (`stat_boxplot()`).
```



The box-plot reflects the frequency data

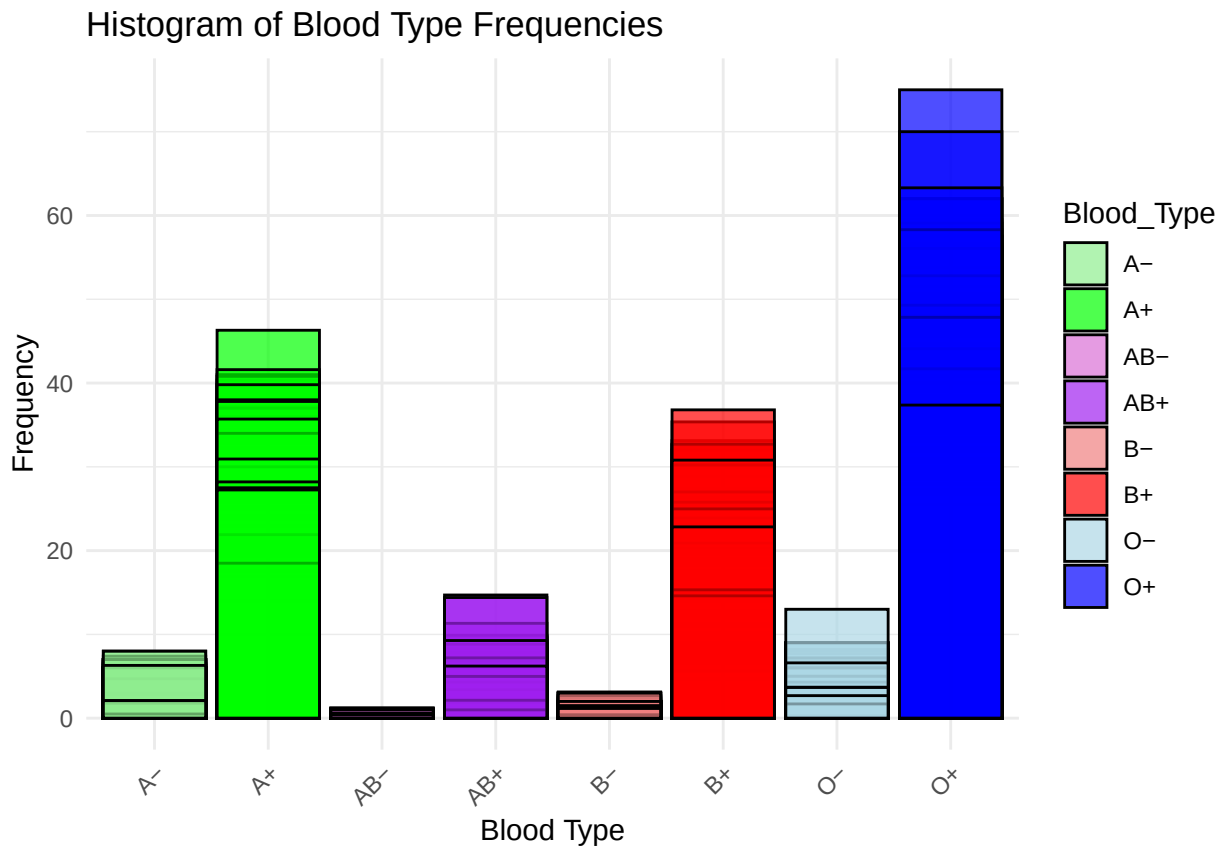
Create a histogram for the blood type frequencies

This code defines a function named create_blood_type_histogram which creates histograms for blood type frequencies, with each blood type represented by a different color.

```
### Create histograms for each blood type with colors (function)
create_blood_type_histogram <- function(blood_type_data) {
  blood_type_data %>%
    pivot_longer(cols = -c(Country, Population), names_to = "Blood_Type", values_to = "Frequency") %>%
    ggplot(aes(x = Blood_Type, y = Frequency, fill = Blood_Type)) +
    geom_bar(stat = "identity", position = "dodge", color = "black", alpha = 0.7) +
    scale_fill_manual(values = c(
      "O+" = "blue",
      "A+" = "green",
      "B+" = "red",
      "AB+" = "purple",
      "O-" = "lightblue",
      "A-" = "lightgreen",
      "B-" = "lightcoral",
      "AB-" = "orchid"
    )) +
    labs(title = "Histogram of Blood Type Frequencies",
         x = "Blood Type",
```

```
         y = "Frequency") +
    theme_minimal() +
    theme(axis.text.x = element_text(angle = 45, hjust = 1))
}

### initialize histogram function
create_blood_type_histogram(blood_type_data)
```

```
## Warning: Removed 4 rows containing missing values (`geom_bar()`).
```



The histogram reflects the frequency data