Exercise 6.2: Histograms, Box Plots, & Bullet Charts

Anuj Tanwar

2/26/2023

Plots Using R

```
knitr::opts_chunk$set(echo = TRUE, warning = FALSE)

# Set Working Directory
setwd("C:/Users/anujt/git/temp/DSC640/Week11&12/ex6-2/")

# Load libraries
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.1.3
```

```
library(stringr)  # for converting to title case
library(reshape2) # for melting data
library(tm)       # for text cleaning
```

```
## Warning: package 'tm' was built under R version 4.1.3

## Loading required package: NLP

## Warning: package 'NLP' was built under R version 4.1.1

##
## Attaching package: 'NLP'

## The following object is masked from 'package:ggplot2':
##
##     annotate
```

```
# library(wordcloud2)
library(wordcloud)
```

```
## Warning: package 'wordcloud' was built under R version 4.1.3

## Loading required package: RColorBrewer
```

```
# Set color to Bellevue purple
color = "#0000FF"
```

Load Data

```
# Load data
birthdf <- read.csv('C:/Users/anujt/git/temp/DSC640/Week11&12/ex6-2/birth-rate.csv')
educadf <- read.csv('C:/Users/anujt/git/temp/DSC640/Week11&12/ex6-2/education.csv')
eduSummary <- read.csv("C:/Users/anujt/git/temp/DSC640/Week11&12/ex6-2/education_summary.csv")
```

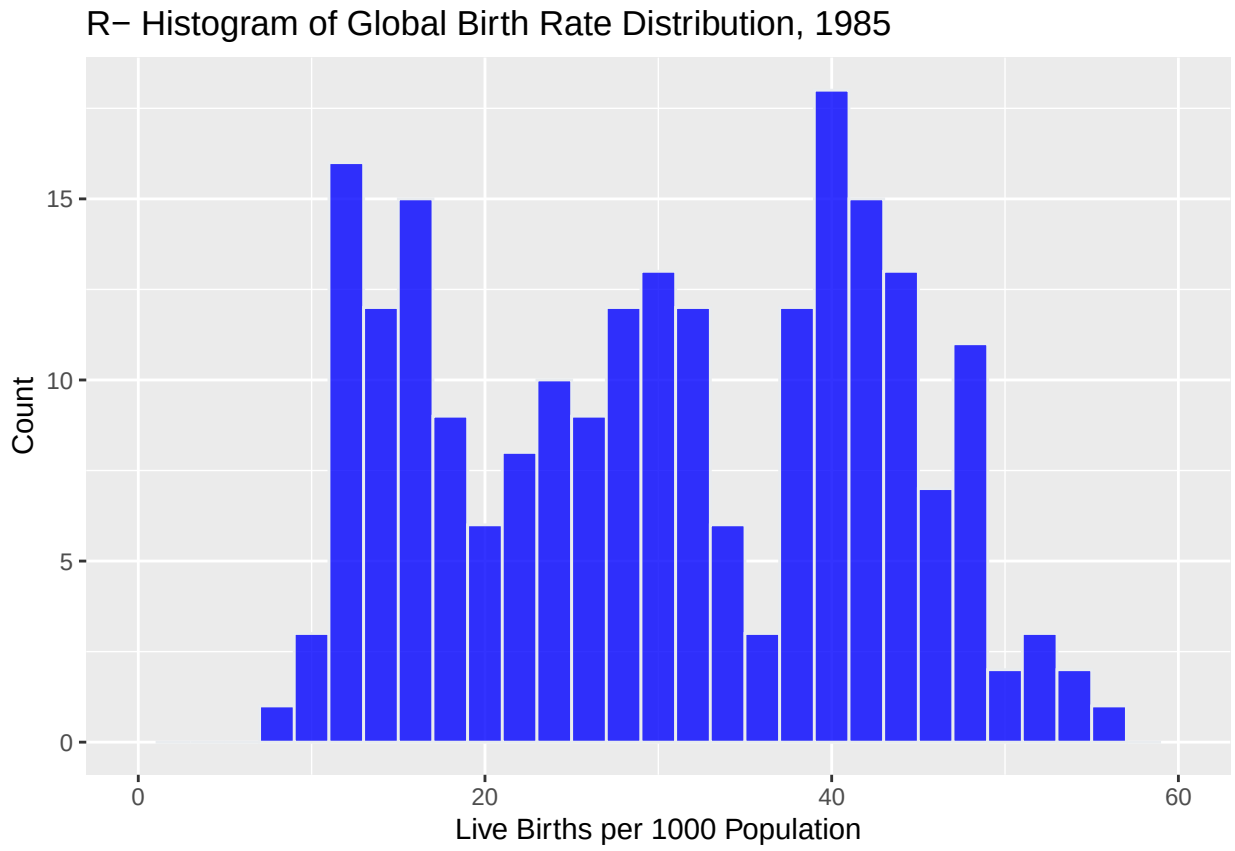
Clean Data

```
# Reshape education data set
edumelt <- melt(educadf[,1:4], id="state")
# Save reformatted education data as CSV for use elsewhere
write.csv(edumelt, "education_pivot.csv", row.names = FALSE)

# Rename first column of summarized education data
names(eduSummary)[1] <- 'Category'
```

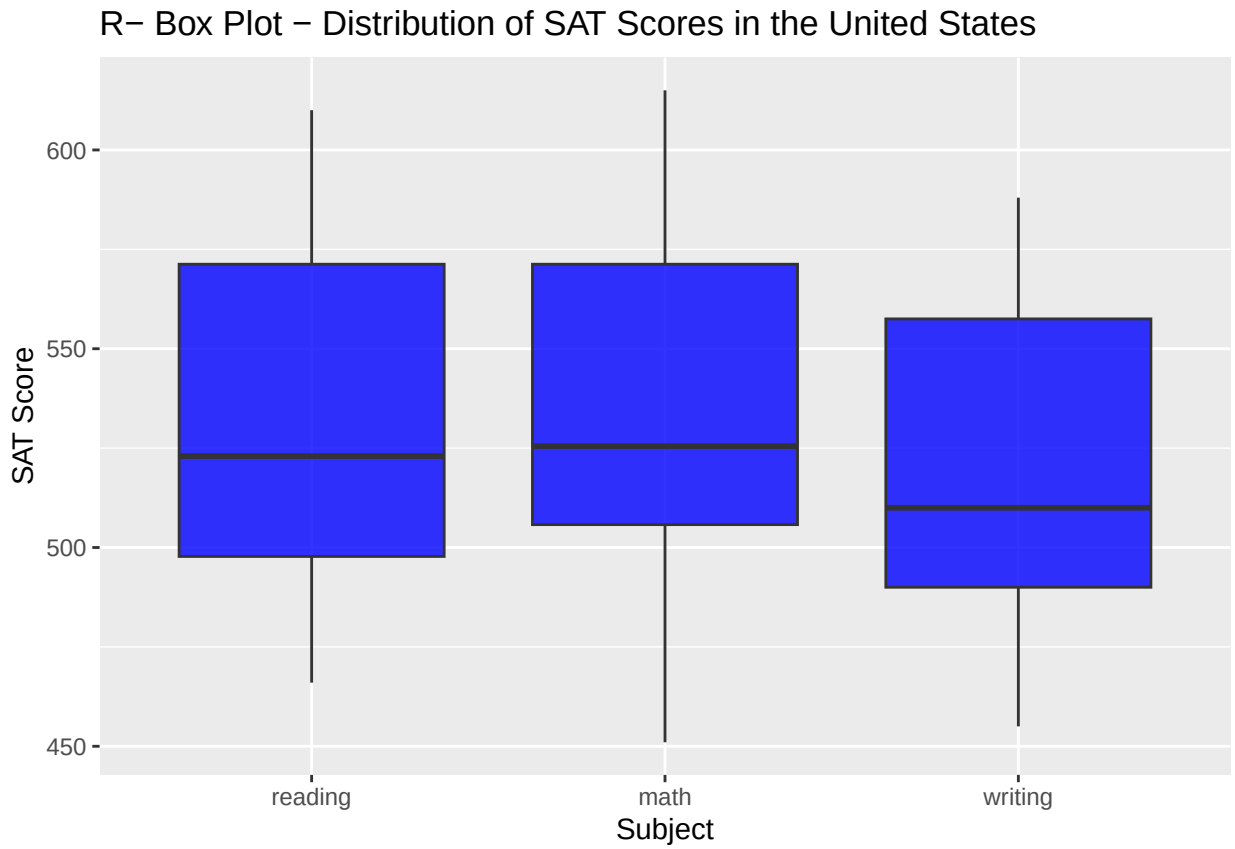
Histogram

```
# Plot histogram
ggplot(birthdf, aes(x=X1985)) +
  geom_histogram(binwidth = 2, fill=color, color="#e9ecef", alpha=0.8) +
  xlim(0,60) +
  ggtitle('R- Histogram of Global Birth Rate Distribution, 1985') +
  labs(x="Live Births per 1000 Population", y="Count")
```



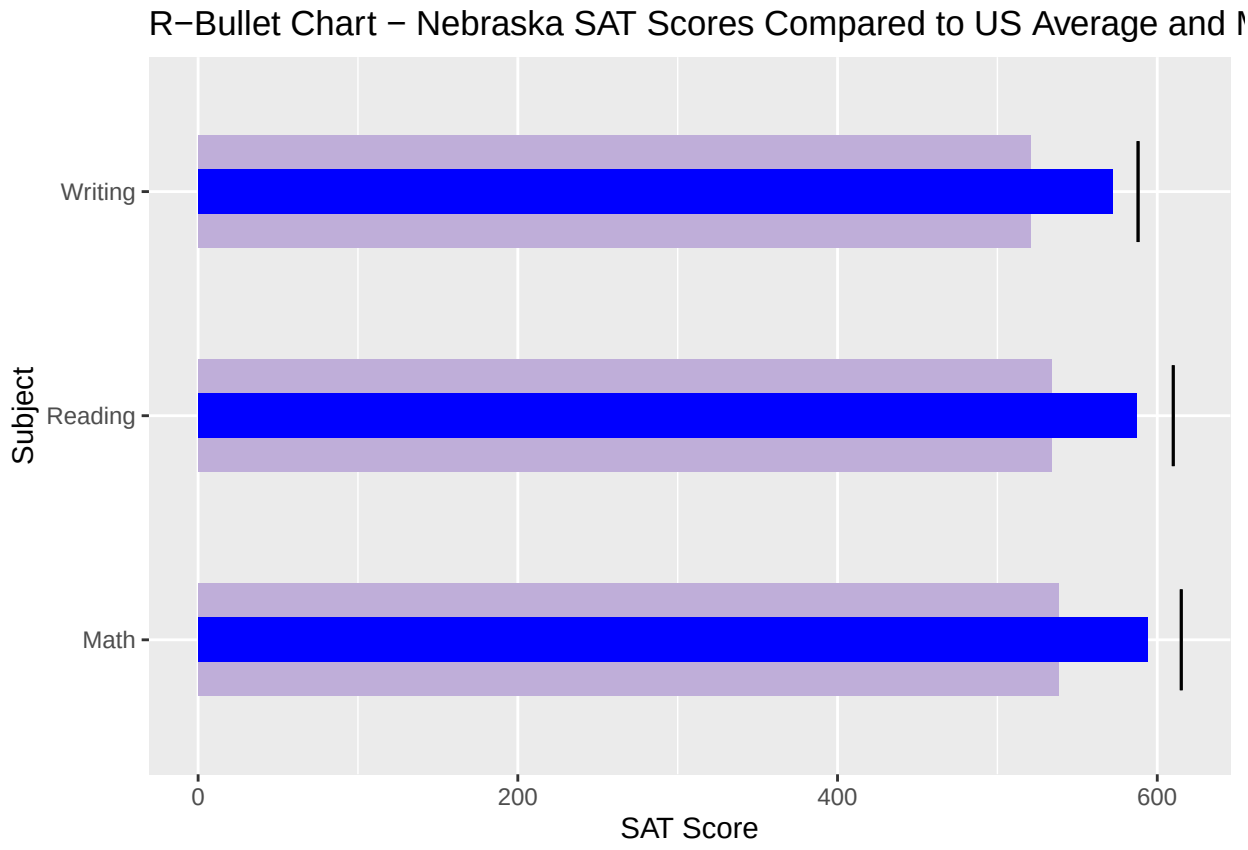
Box Plot

```
# Make box & whisker plot
ggplot(edumelt, aes(x=variable, y=value)) +
  geom_boxplot(fill=color, alpha=0.8) +
  ggtitle('R- Box Plot - Distribution of SAT Scores in the United States') +
  labs(x="Subject", y="SAT Score")
```



Bullet Chart

```
# Create bullet chart
ggplot(eduSummary, aes(Category, Average)) +
  geom_col(fill="#bfaed9", width = 0.5) +
  geom_col(fill=color, aes(Category, Actual), width = 0.2) +
  geom_errorbar(aes(y = Max, x = Category,
                    ymin = Max, ymax = Max),
                width = 0.45) +
  coord_flip() +
  ggtitle('R-Bullet Chart - Nebraska SAT Scores Compared to US Average and Max Score') +
  labs(x="Subject", y="SAT Score")
```



BYO Chart: Word Cloud

```
# Load text data
text <- read.csv("C:/Users/anujt/git/temp/DSC640/Week11&12/ex6-2/compiled_Text.txt", sep = "\t", header
# Create corpus
corp <- VCorpus(VectorSource(text))

# Clean up text data
corp <- tm_map(corp, removeNumbers)
corp <- tm_map(corp, removePunctuation)
corp <- tm_map(corp, stripWhitespace)
corp <- tm_map(corp, content_transformer(tolower))
corp <- tm_map(corp, removeWords, stopwords("english"))

# Create a document-term-matrix
dtm <- TermDocumentMatrix(corp)
matrix <- as.matrix(dtm)
words <- sort(rowSums(matrix), decreasing = TRUE)
df <- data.frame(word = names(words), freq=words)

# Generate word cloud
layout(matrix(c(1, 2), nrow=2), heights=c(1, 4))
par(mar=rep(0, 4))
plot.new()
text(x=0.5, y=0.5, "R - Word Cloud wikipedia Data Science Definition")
wordcloud(words = df$word, freq = df$freq, min.freq = 1,
          max.words = 200, random.order = FALSE,
          colors = brewer.pal(20, "Dark2"))
```

R − Word Cloud wikipedia Data Science Definition