Exercise 6.2: Histograms, Box Plots, & Bullet Charts

Scott Breitbach

2/26/2022

Plots Using R

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

# Set Working Directory
setwd("C:/Users/micha/OneDrive/Documents/GitHub/DSC640/Weeks11-12/")

# Load libraries
library(ggplot2)
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 4.1.2
```

```
## Loading required package: NLP
```

```
##
## Attaching package: 'NLP'
```

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

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

```
## Loading required package: RColorBrewer
```

```
# Set color to Bellevue purple
color = "#4f3674"
```

Load Data

```r
# Load data
birthdf <- read.csv('birth-rate.csv')
educadf <- read.csv('education.csv')
eduSummary <- read.csv("education_summary.csv")
```

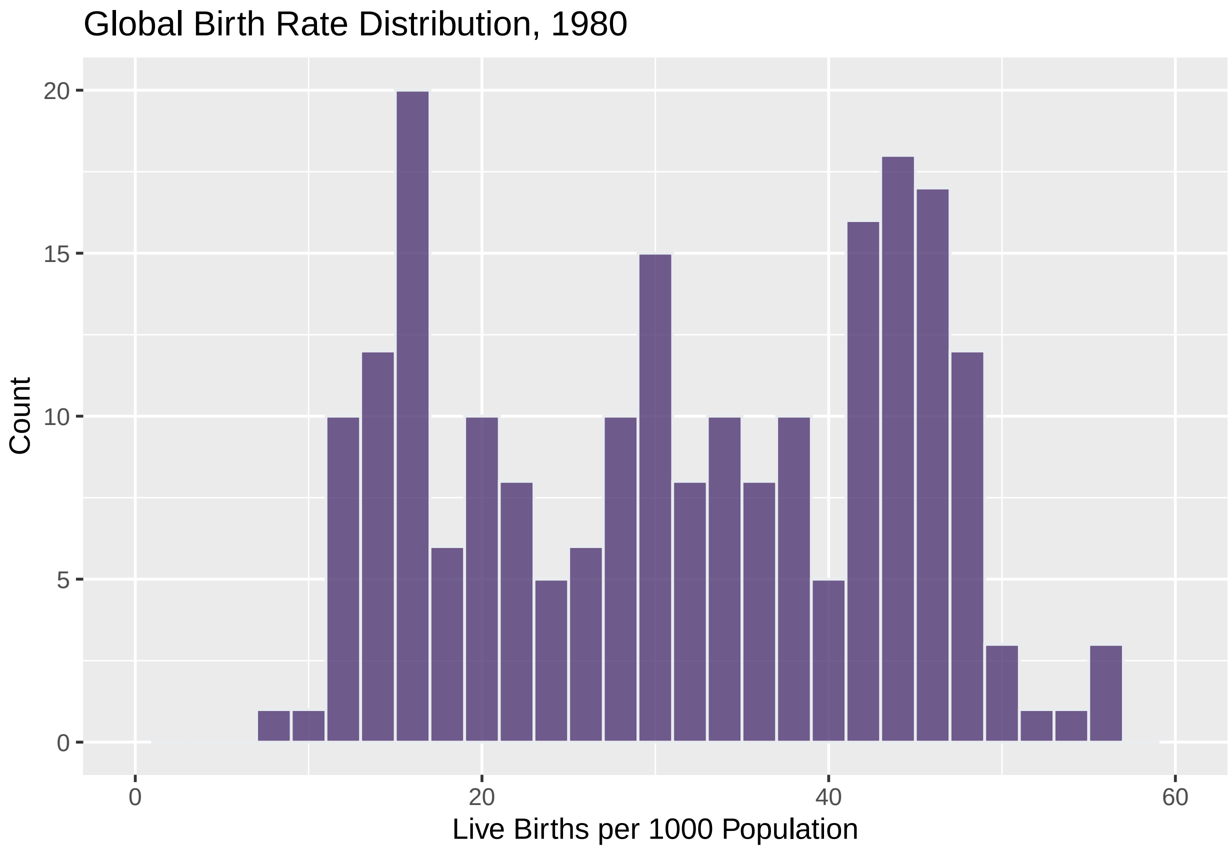
Clean Data

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

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

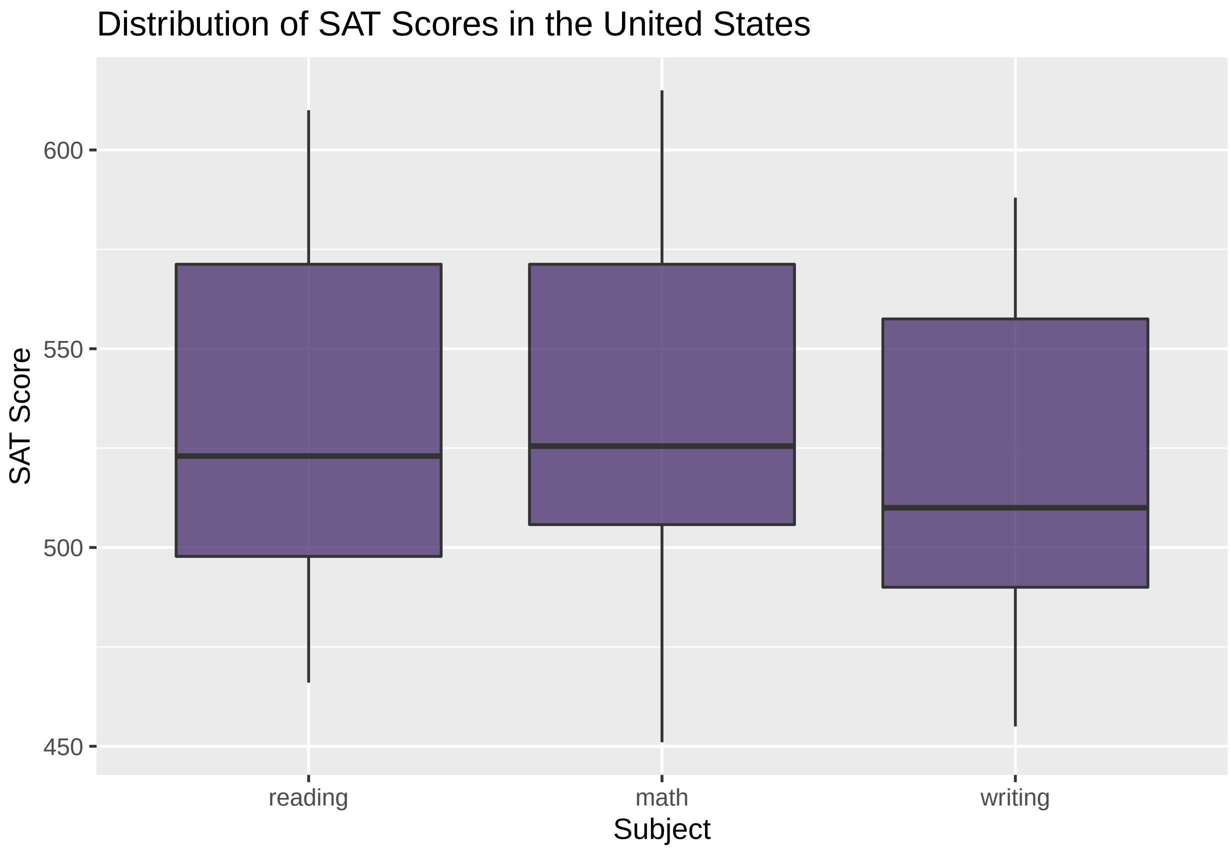
Histogram

```
# Plot histogram
ggplot(birthdf, aes(x=X1980)) +
  geom_histogram(binwidth = 2, fill=color, color="#e9ecef", alpha=0.8) +
  xlim(0,60) +
  ggtitle('Global Birth Rate Distribution, 1980') +
  labs(x="Live Births per 1000 Population", y="Count")
```



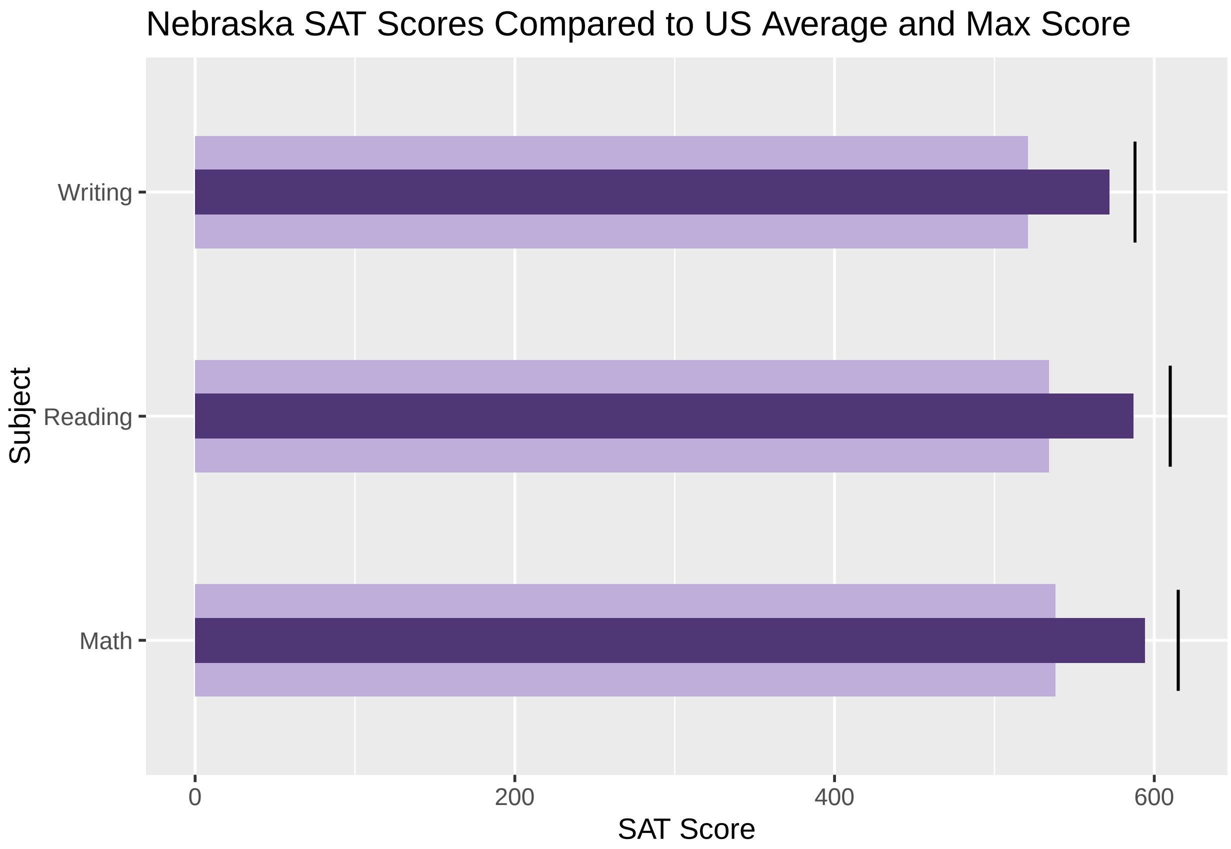
Box Plot

```
# Make box & whisker plot
ggplot(edumelt, aes(x=variable, y=value)) +
  geom_boxplot(fill=color, alpha=0.8) +
  ggtitle('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('Nebraska SAT Scores Compared to US Average and Max Score') +
  labs(x="Subject", y="SAT Score")
```



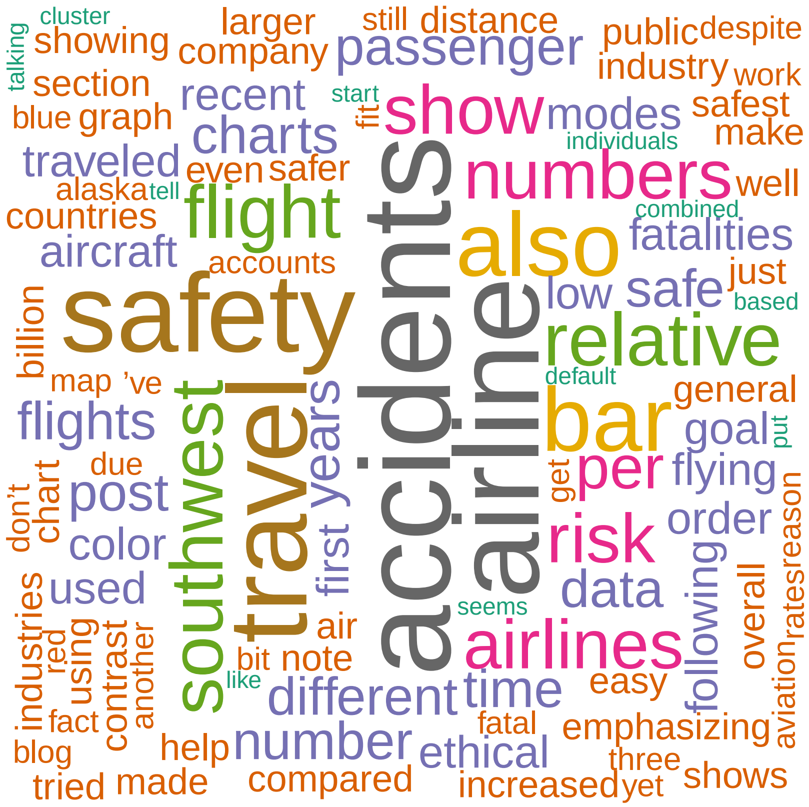
BYO Chart: Word Cloud

```r
# Load text data
text <- read.csv("compiled_words.txt", sep = "\t", header = FALSE)
# 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
wordcloud(words = df$word, freq = df$freq, min.freq = 1,
          max.words = 200, random.order = FALSE,
          colors = brewer.pal(20, "Dark2"))
```


talking
cluster
showing
larger
company
still
distance
passenger
public
despite
industry
work
section
recent
start
fit
show
modes
safest
blue
graph
charts
individuals
make
traveled
even
safer
numbers
well
alaska
tell
flight
combined
countries
fatalities
aircraft
accounts
also
just
billion
safety
low
safe
based
relative
accidents
airline
default
map
've
general
flights
bar
goal
put
travel
years
get
per
flying
southwest
due
don't
chart
post
reason
order
first
risk
color
following
overall
rates
used
data
seems
industries
red
using
contrast
another
air
airlines
aviation
bit
note
like
easy
fact
different
time
fatal
emphasizing
blog
help
number
ethical
three
shows
tried
made
compared
increased
yet