Exercise 6.2: Histograms, Box Plots, & Bullet Charts

Scott Breitbach

DSC640 - 02/26/2022

Plots Using **Python**

Load Data

In [14]:

```
# Load libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from wordcloud import WordCloud, STOPWORDS, ImageColorGenerator
```

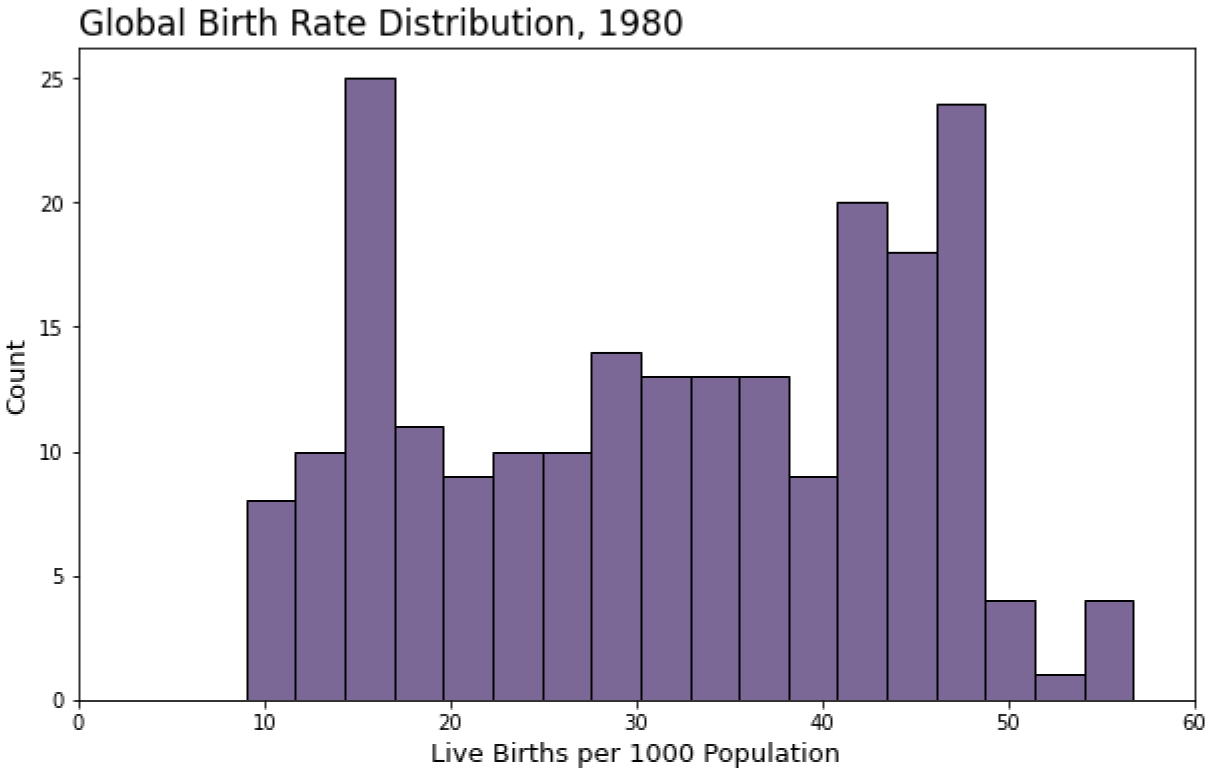
In [15]:

```
# Load data
birthDF = pd.read_csv("birth-rate.csv")
educaDF = pd.read_csv("education.csv")
edumelt = pd.read_csv("education_melted.csv")
scoresNE = pd.read_csv("education_summary.csv")
textDF = pd.read_csv("clean_text.csv", encoding='cp1252')

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

Histogram

In [3]:

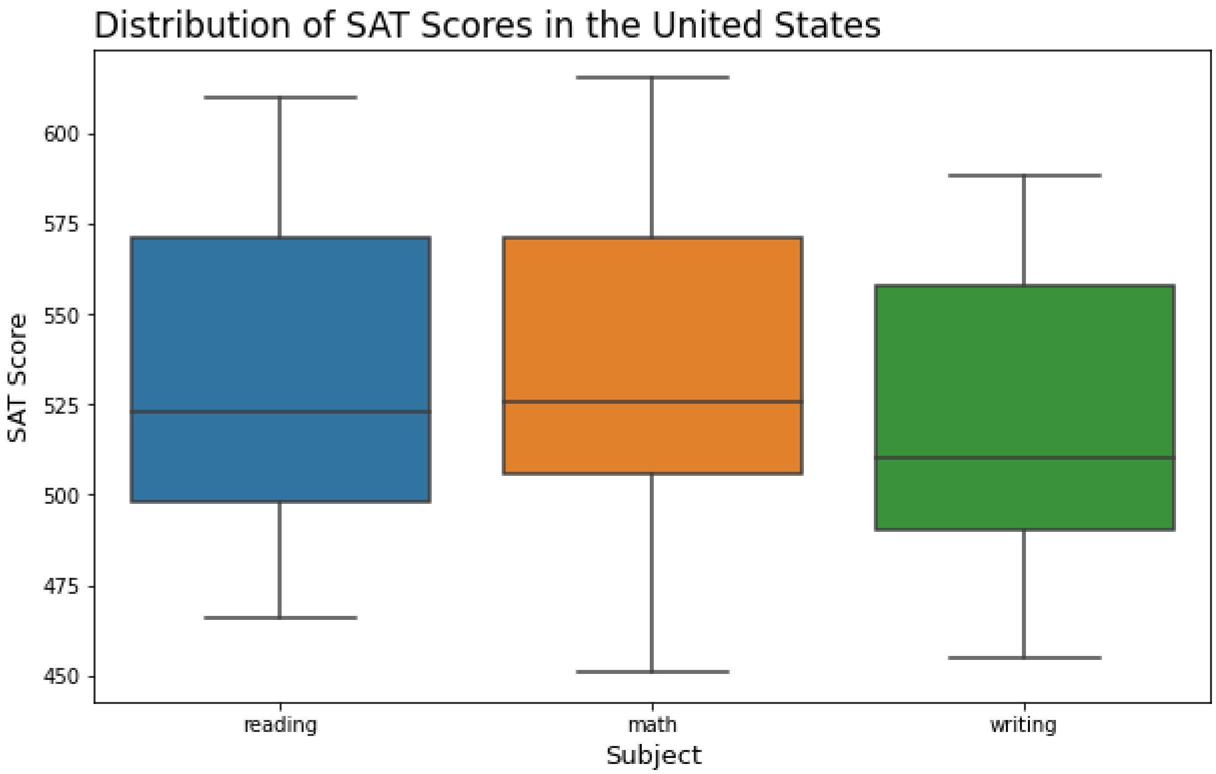
```
# Initialize the matplotlib figure
f, ax = plt.subplots(figsize=(10, 6))

# Plot histogram
sns.histplot(birthDF, x="1980", bins=18, color=color)

# Add chart title and labels
plt.title("Global Birth Rate Distribution, 1980",
          fontsize = 17, loc = 'left')
plt.xlabel("Live Births per 1000 Population", fontsize = 13)
plt.ylabel("Count", fontsize = 13)
plt.xlim(0,60)  # Set min and max for x-axis

plt.show()
```



Box Plot

In [4]:

```
# Initialize the matplotlib figure
f, ax = plt.subplots(figsize=(10, 6))

# Create box plot
sns.boxplot(x=edumelt['variable'], y=edumelt['value'])

# Add chart title and labels
plt.title("Distribution of SAT Scores in the United States",
          fontsize = 17, loc = 'left')
plt.xlabel("Subject", fontsize = 13)
plt.ylabel("SAT Score", fontsize = 13)

plt.show()
```



Bullet Chart

In [9]:

```
# Prepare data for graphing
lims = [scoresNE.iloc[0,4], scoresNE.iloc[0,2], scoresNE.iloc[0,8]]
data_to_plt = (scoresNE.iloc[0,0], scoresNE.iloc[0,1], scoresNE.iloc[0,6])
```

In [10]:

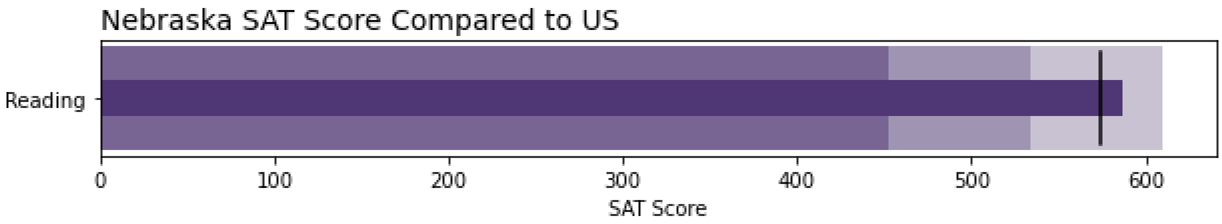
```
# Build a color palette
palette = sns.light_palette(color, len(lims)+2, reverse=True)
```

In [13]:

```
# Build the stacked bar chart of the ranges
fig, ax = plt.subplots(figsize=(10,3))
ax.set_aspect('equal')
ax.set_yticks([1])
ax.set_yticklabels([data_to_plt[0]])

prev_limit = 0
for idx, lim in enumerate(lims):
    ax.barh([1], lim-prev_limit, left=prev_limit, height=60, color=palette[idx+1])
    prev_limit = lim

# Draw the value we're measuring
ax.barh([1], data_to_plt[1], color=palette[0], height=20)

# Add the target marker
ax.axvline(data_to_plt[2], color="black", ymin=0.10, ymax=0.9)

# Add title and labels
plt.title("Nebraska SAT Score Compared to US",
          fontsize = 14, loc = 'left')
# fig.suptitle("Nebraska SAT Score Compared to US", fontsize=14)
ax.set_xlabel("SAT Score")
```

Out[13]:

```
Text(0.5, 0, 'SAT Score')
```



BYO Chart: Word Cloud

In [16]:

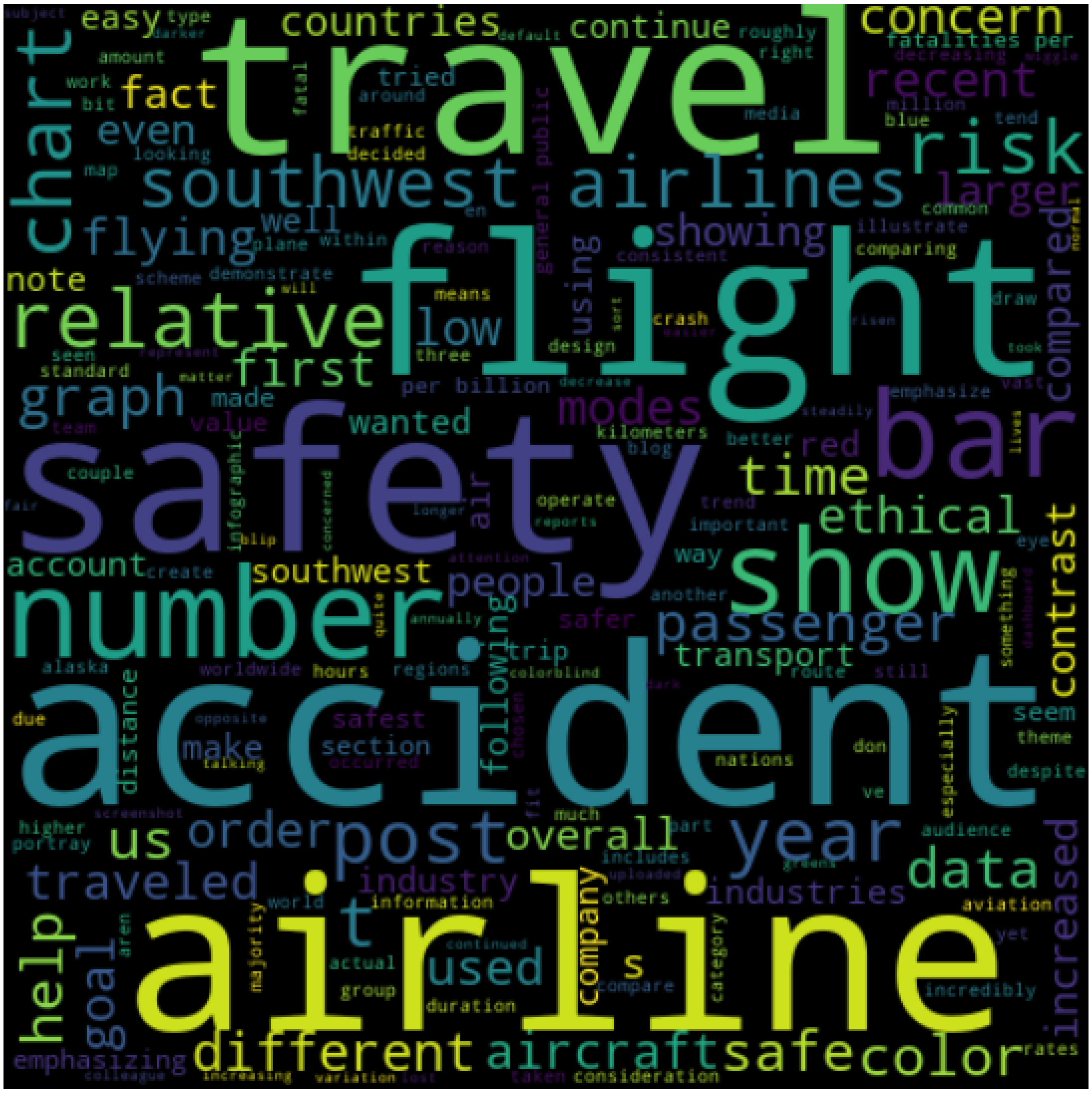
```
# Concatenate text
text = ''
for i in range(len(textDF['x'])):
    text += textDF['x'][i]
```

In [18]:

```
# Create the wordcloud object
wordcloud = WordCloud(width=480, height=480, margin=0).generate(text)
```

In [19]:

```
# Initialize the matplotlib figure
f, ax = plt.subplots(figsize=(10, 10))

# Display the generated image:
plt.imshow(wordcloud, interpolation='bilinear')
plt.axis("off")
plt.margins(x=0, y=0)
plt.show()
```



In [ ]:

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

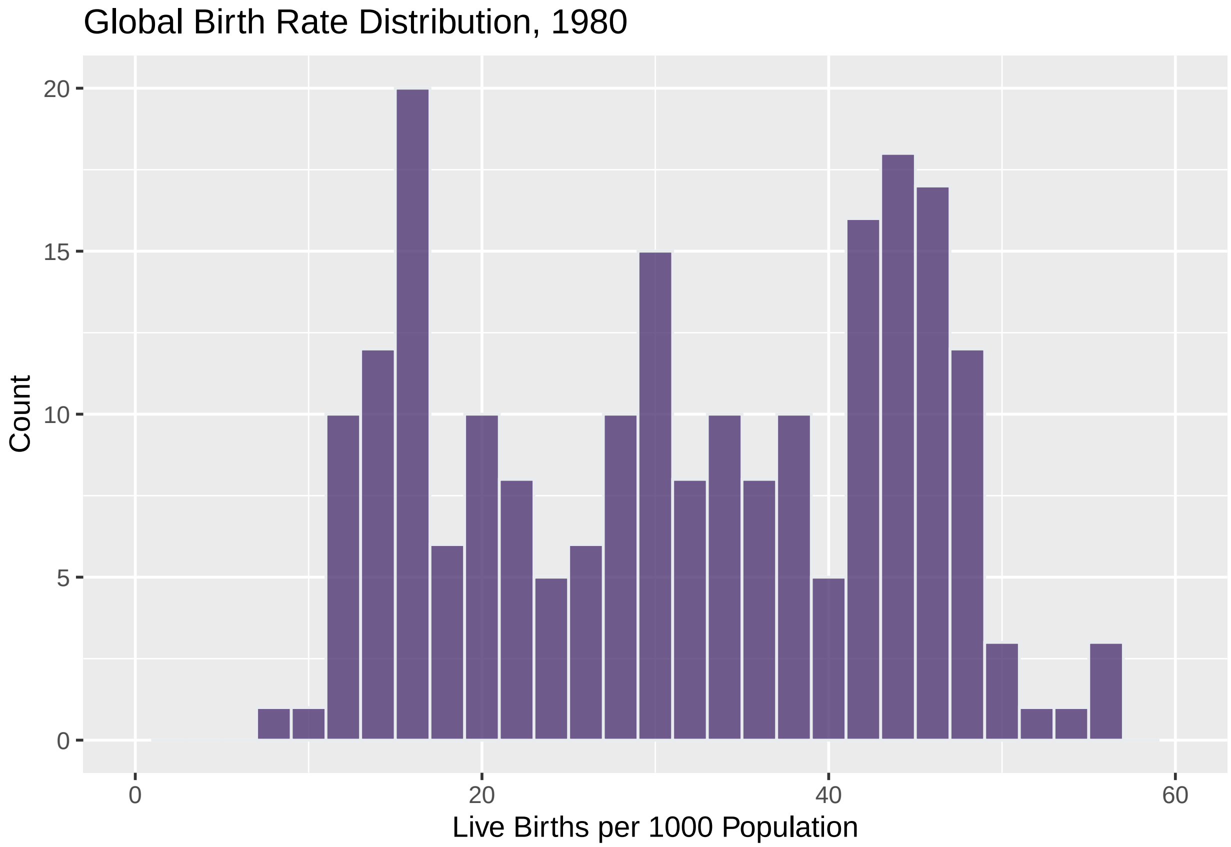
```r
# Load data
birthdf <- read.csv('birth-rate.csv')
educadf <- read.csv('education.csv')
eduSummary <- read.csv("education_summary.csv")
```

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'
```

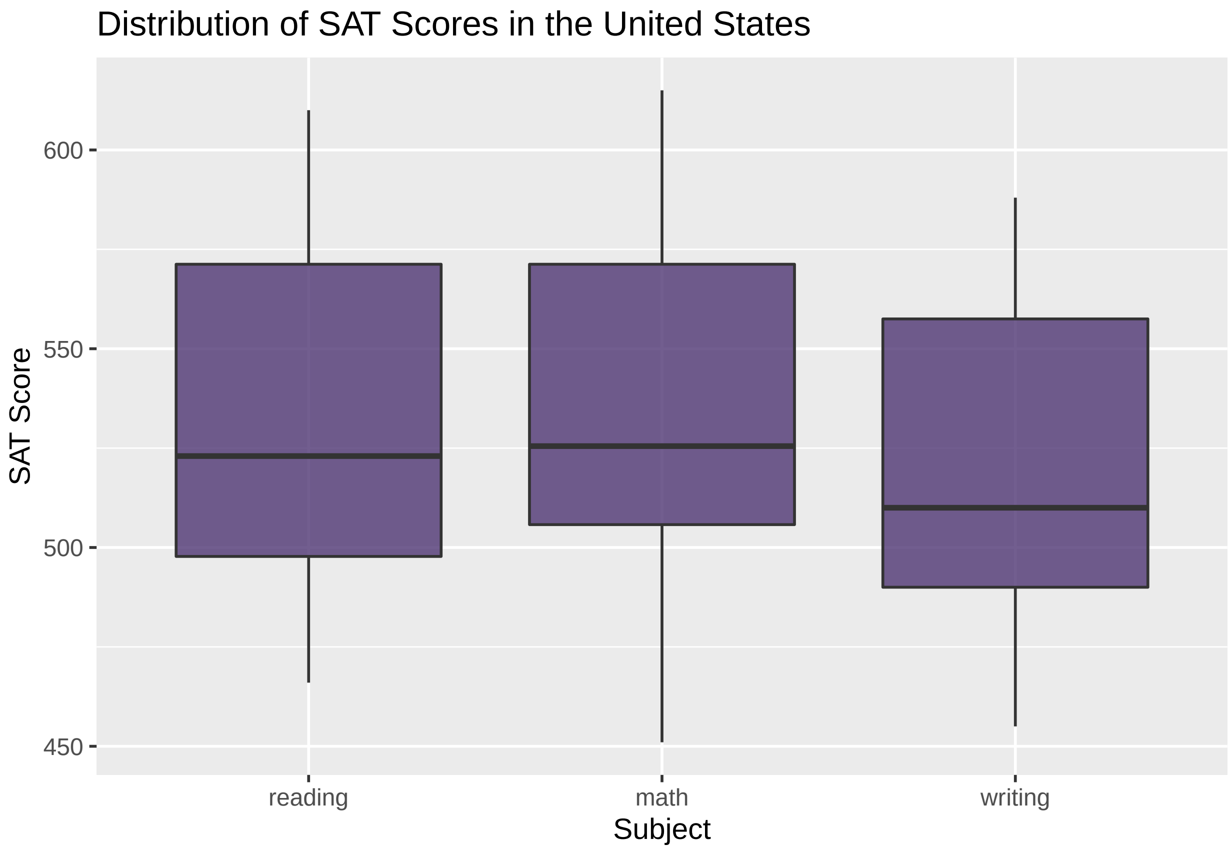
Histogram

```
# 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")
```



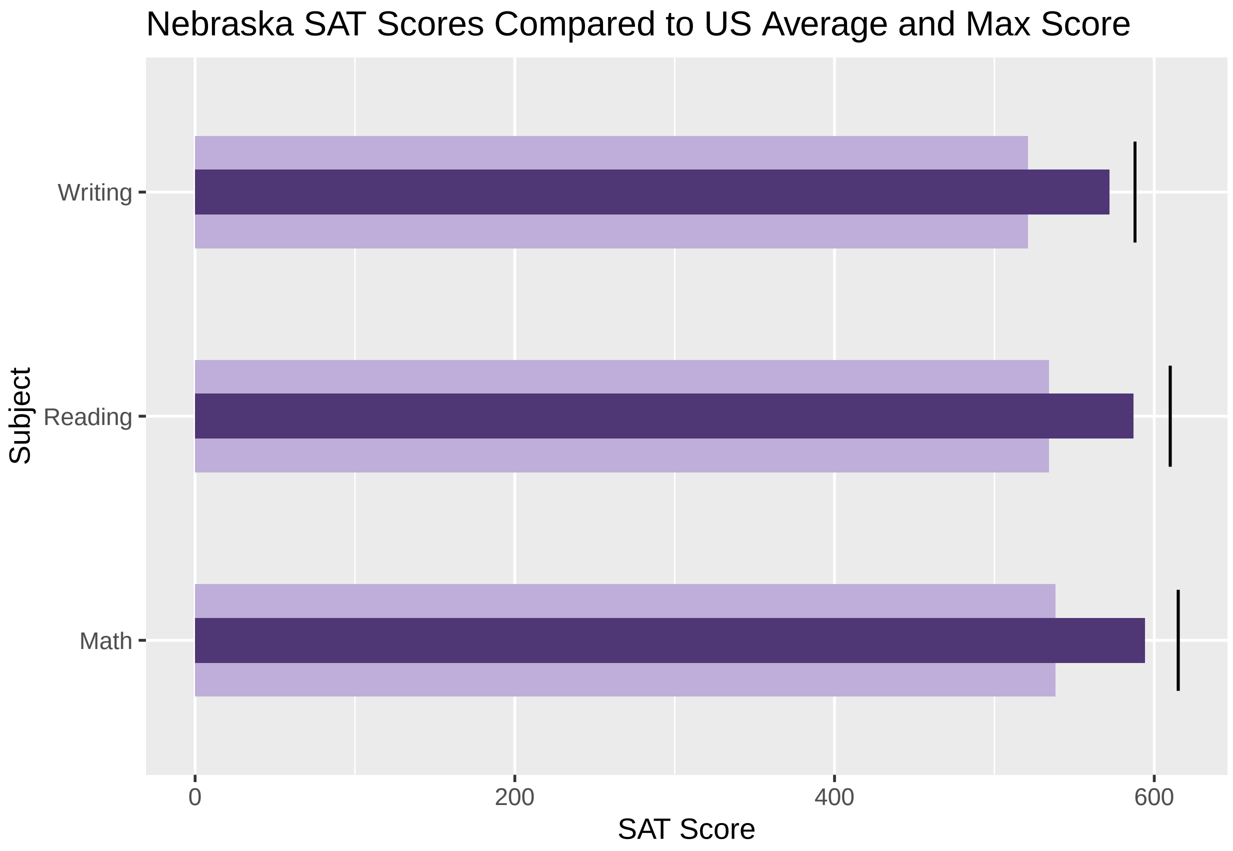
Box Plot

```r
# 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")
```



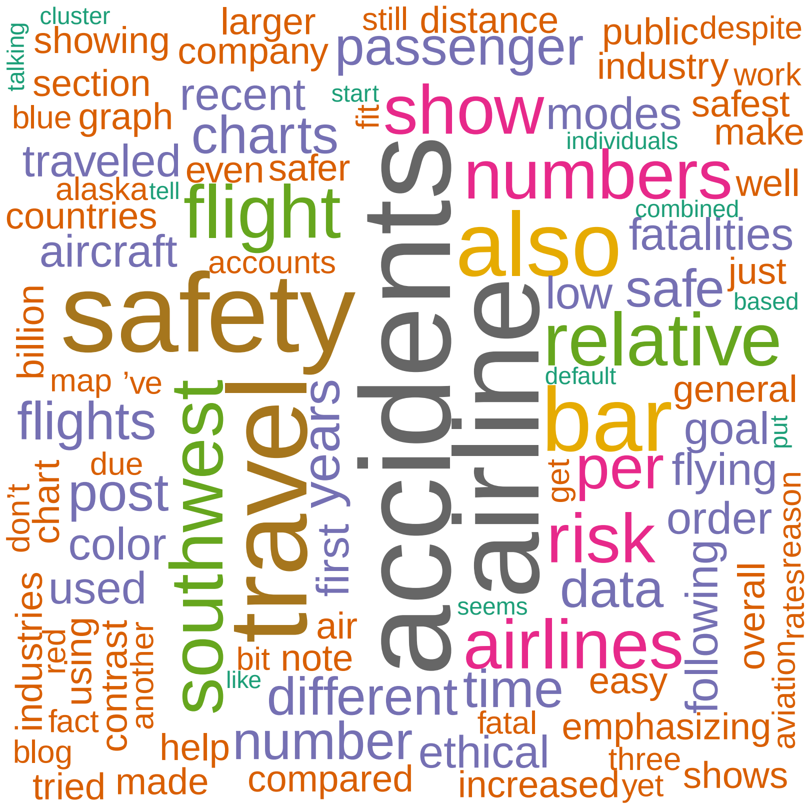
Bullet Chart

```
# 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

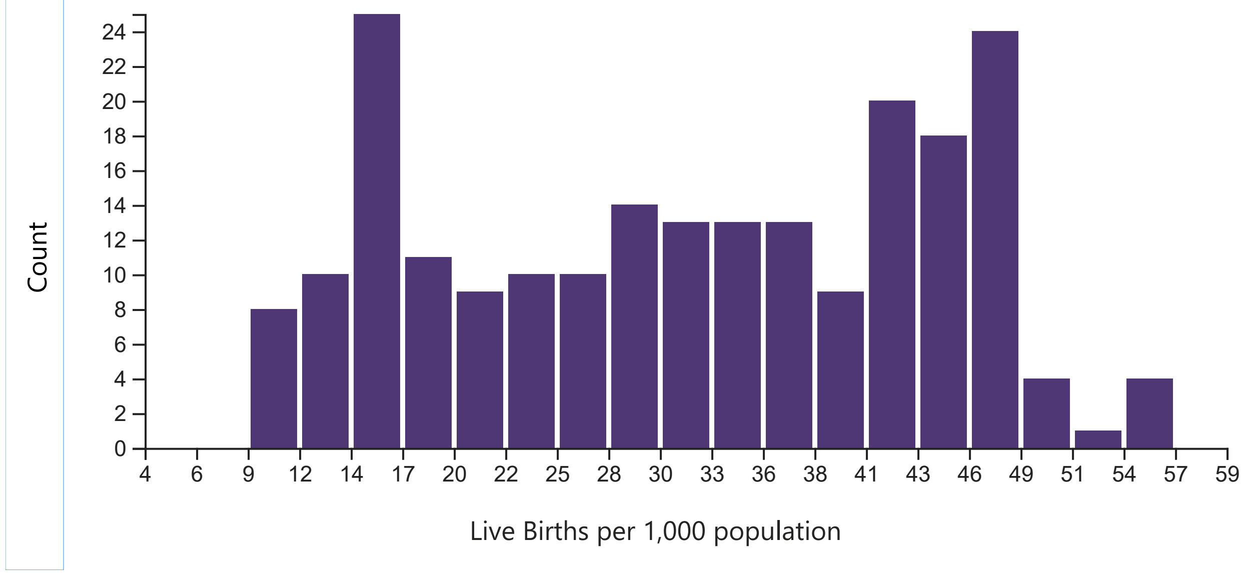
```
# Load text data
text <- read.csv("compiled_words.txt", sep = "\t", header = FALSE)
# Create corpus
corp <- VCorpus(VectorSource(text))

# 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"))
```



Power BI Charts

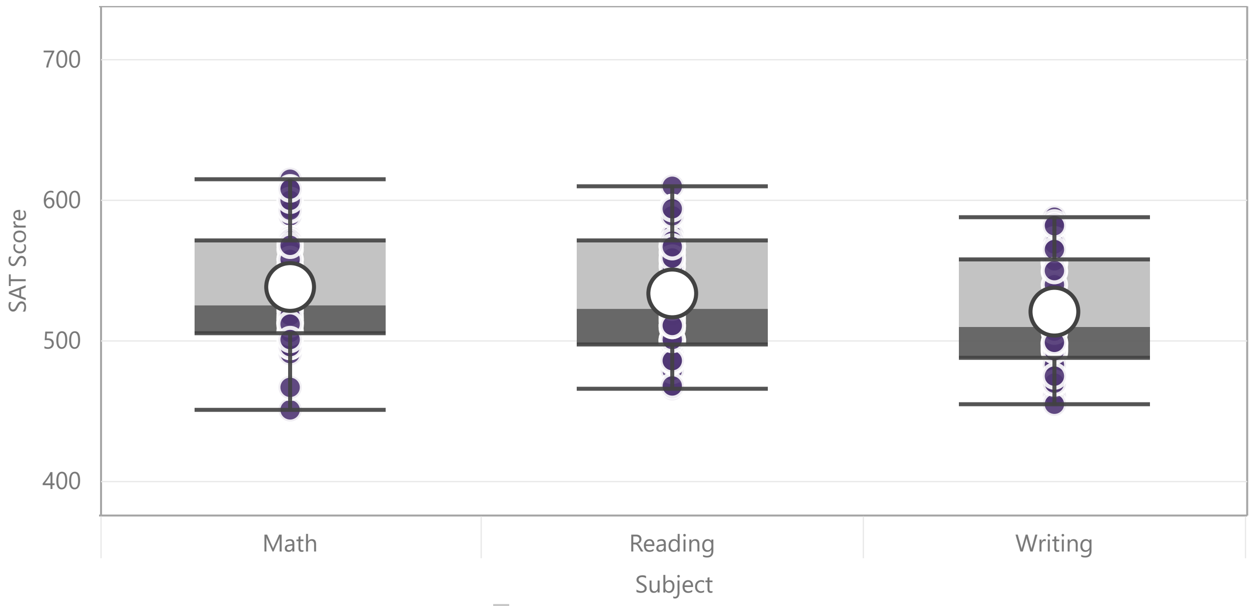
Histogram

Global Birth Rate Distribution, 1980



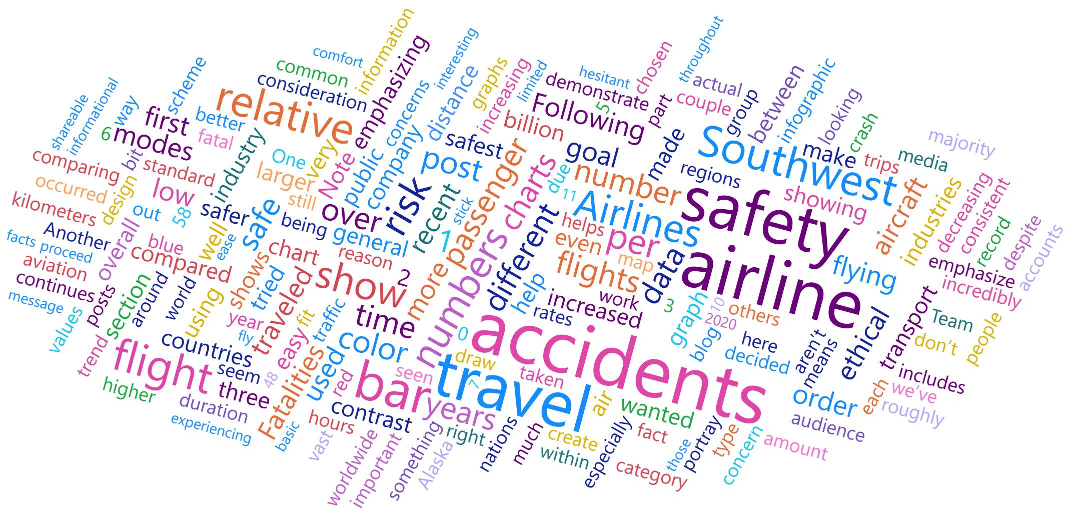
Box & Whisker Plot

Distribution of AT Scores in the United States



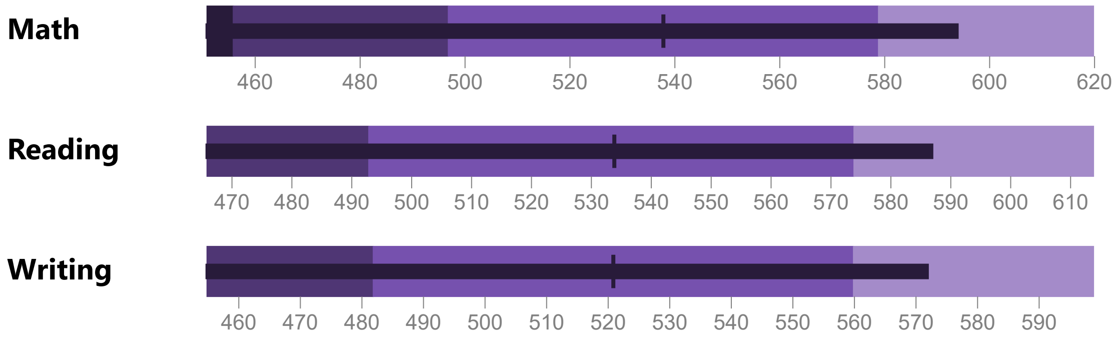
Word Cloud

Words Used in My Project Assignment Papers



Bullet Chart

Nebraska SAT Scores Compared to US Average


Math
Reading
Writing