Young_DSC 640_Assignment 6.2_R

Bret Young

11/12/2020

```
# load data
file_1 = 'birth-rate.csv'
data_1 = read.delim(file_1, header = TRUE, sep = ',')
```

```
# Print data
head(data_1)
```

Country <fctr>	X1960 <dbl>	X1961 <dbl>	X1962 <dbl>	X1963 <dbl>	X1964 <dbl>	X1965 <dbl>	X1966 <dbl>	X1967 <dbl>
1 Aruba	36.40000	35.179	33.863	32.459	30.994	29.51300	28.069	26.721
2 Afghanistan	52.20100	52.206	52.208	52.204	52.192	52.16800	52.130	52.076
3 Angola	54.43200	54.394	54.317	54.199	54.040	53.83600	53.585	53.296
4 Albania	40.88600	40.312	39.604	38.792	37.913	37.00800	36.112	35.245
5 Netherlands Antilles	32.32100	30.987	29.618	28.229	26.849	25.51800	24.280	23.173
6 Arab World	47.61122	*NA*	*NA*	*NA*	*NA*	46.57288	*NA*	*NA*

6 rows | 1-10 of 51 columns

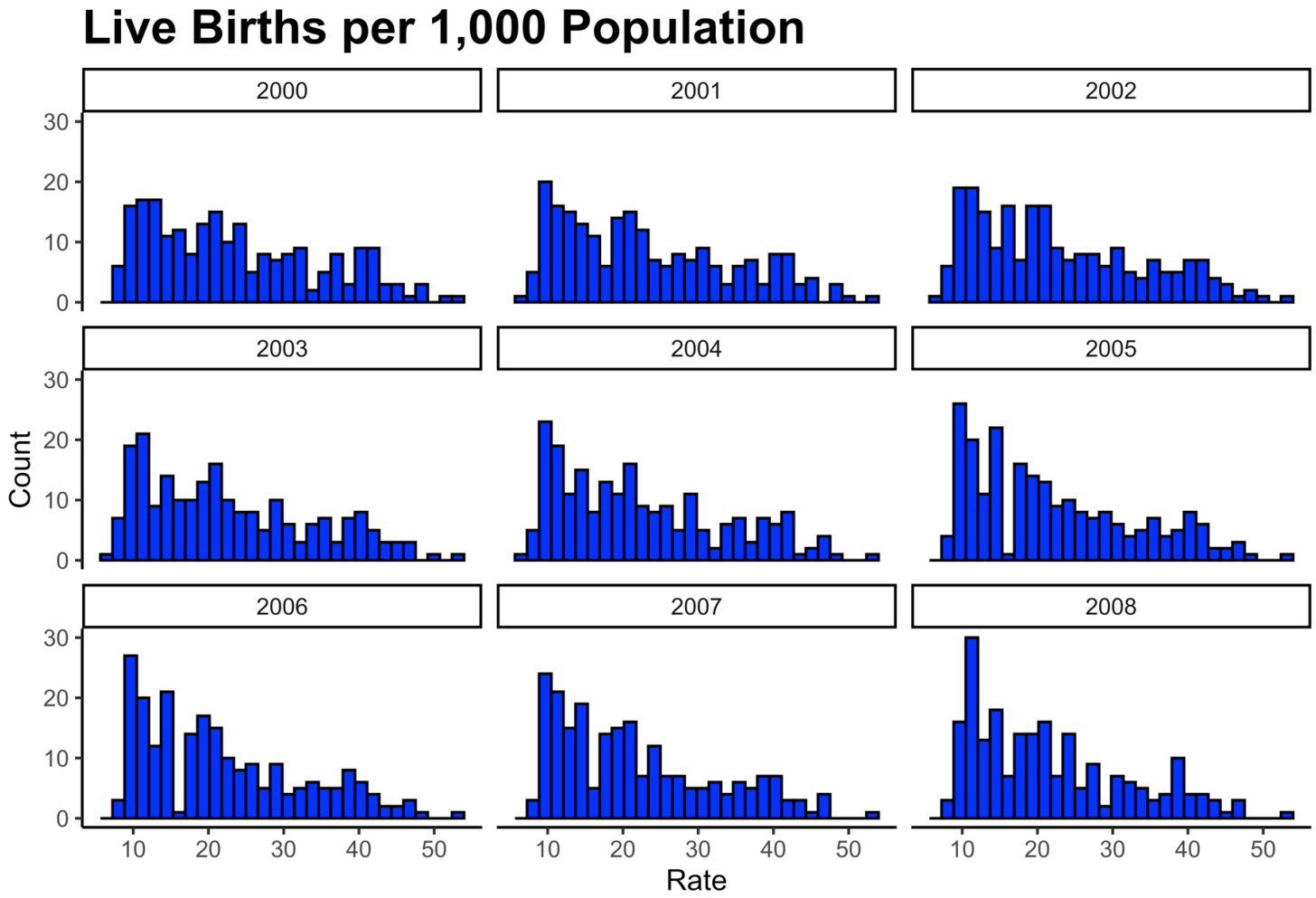
```
# load library
library(tidyverse)

# Select 2000 thru 2008
data_1_filtered = data_1[,(ncol(data_1)-8):ncol(data_1)]

# convert data to long form
data_1_filtered = gather(data_1_filtered)

# remove 'X' from year names
data_1_filtered = data_1_filtered %>% mutate(key = gsub("X", "", key))

head(data_1_filtered)
```

key	value

	<chr>	<dbl>
1	2000	14.5280
2	2000	50.9030
3	2000	48.3550
4	2000	16.8500
5	2000	15.4120
6	2000	28.6828

6 rows

```
# load library
library(ggplot2)

# create histograms for 2000 thru 2008
ggplot(data_1_filtered, aes(value)) +
  geom_histogram(fill="blue", color="black") +
  facet_wrap(~key) +
  ggtitle("Live Births per 1,000 Population") +
  labs(caption = "Source: Data Collected by Nathan Yau from The World Bank",
      x = "Rate",
      y = "Count") +
  theme_classic() +
  theme(plot.title = element_text(face = "bold", size = 18),
        plot.subtitle = element_text(color = "light gray"),
        plot.caption = element_text(color = "light gray")
        )
```



Source: Data Collected by Nathan Yau from The World Bank

```
# load data
file_2 = 'education.csv'
data_2 = read.delim(file_2, header = TRUE, sep = ',')
```

```
# Print data
head(data_2)
```

state <fctr>	reading <int>	... <int>	writing <int>	percent_graduates_sat <int>	pupil_staff_ratio <dbl>	drop
1 United States	501	515	493	46	7.9	
2 Alabama	557	552	549	7	6.7	
3 Alaska	520	516	492	46	7.9	
4 Arizona	516	521	497	26	10.4	
5 Arkansas	572	572	556	5	6.8	

6	California	500	513	498	49	10.9

6 rows

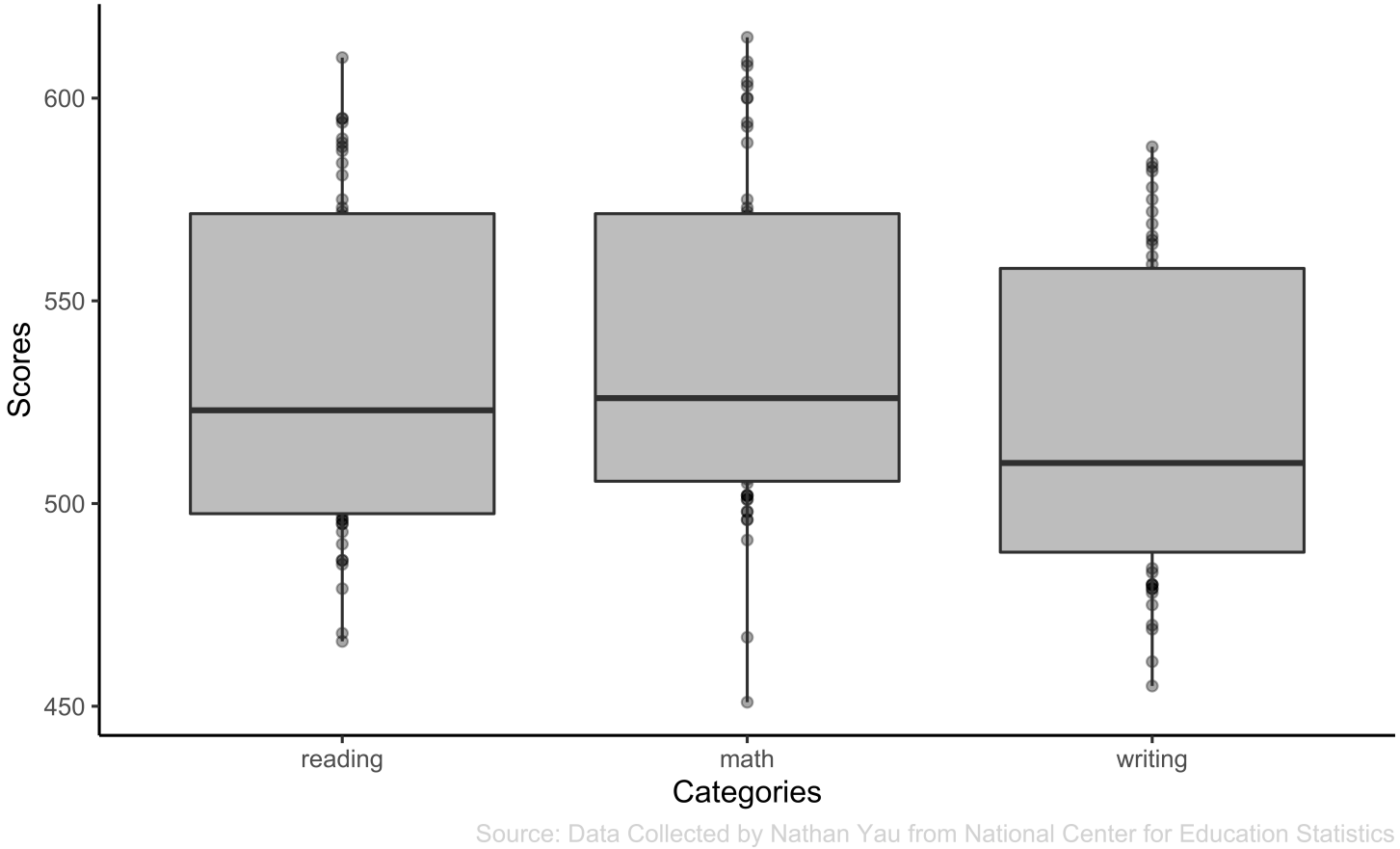
```
# load library
library(reshape2)
library(dplyr)

# filter data to columns needed and remove US
data_2_filtered = data_2[, 1:4] %>% filter(!grepl('United States', state))

# melt data to make three boxplots
data_2_melt = melt(data_2_filtered, id.var = "state")
```

```
head(data_2_melt)
```

	state <fctr>	variable <fctr>	value <int>
1	Alabama	reading	557
2	Alaska	reading	520
3	Arizona	reading	516
4	Arkansas	reading	572
5	California	reading	500
6	Colorado	reading	568

6 rows

```
# create boxplots
ggplot(data_2_melt, aes(x = variable, y = value)) +
  geom_point(alpha = 0.4) +
  geom_boxplot(fill="gray") +
  ggtitle("SAT Category Scores in the United States") +
  labs(caption = "Source: Data Collected by Nathan Yau from National Center for Educa
tion Statistics",
       subtitle = "Reading and Math median scores, 523 and 525 respectively,\nare sim
ilar while Writting scores are lower with a median of 510.",
       x = "Categories",
       y = "Scores") +
  theme_classic() +
  theme(plot.title = element_text(face = "bold", size = 18),
        plot.subtitle = element_text(color = "light gray"),
        plot.caption = element_text(color = "light gray")
        )
```



```
data_2_sat = data_2 %>% filter(state == 'United States')

data_2_sat['Total_SAT_Score'] = round((data_2_sat['math'] + data_2_sat['reading'] + d
ata_2_sat['writing']) * (2/3), 0)

# set SAT percentile values
data_2_sat['sat_49'] = 1050
data_2_sat['sat_50'] = 1070
data_2_sat['sat_74'] = 1200
data_2_sat['sat_91'] = 1350

head(data_2_sat)
```

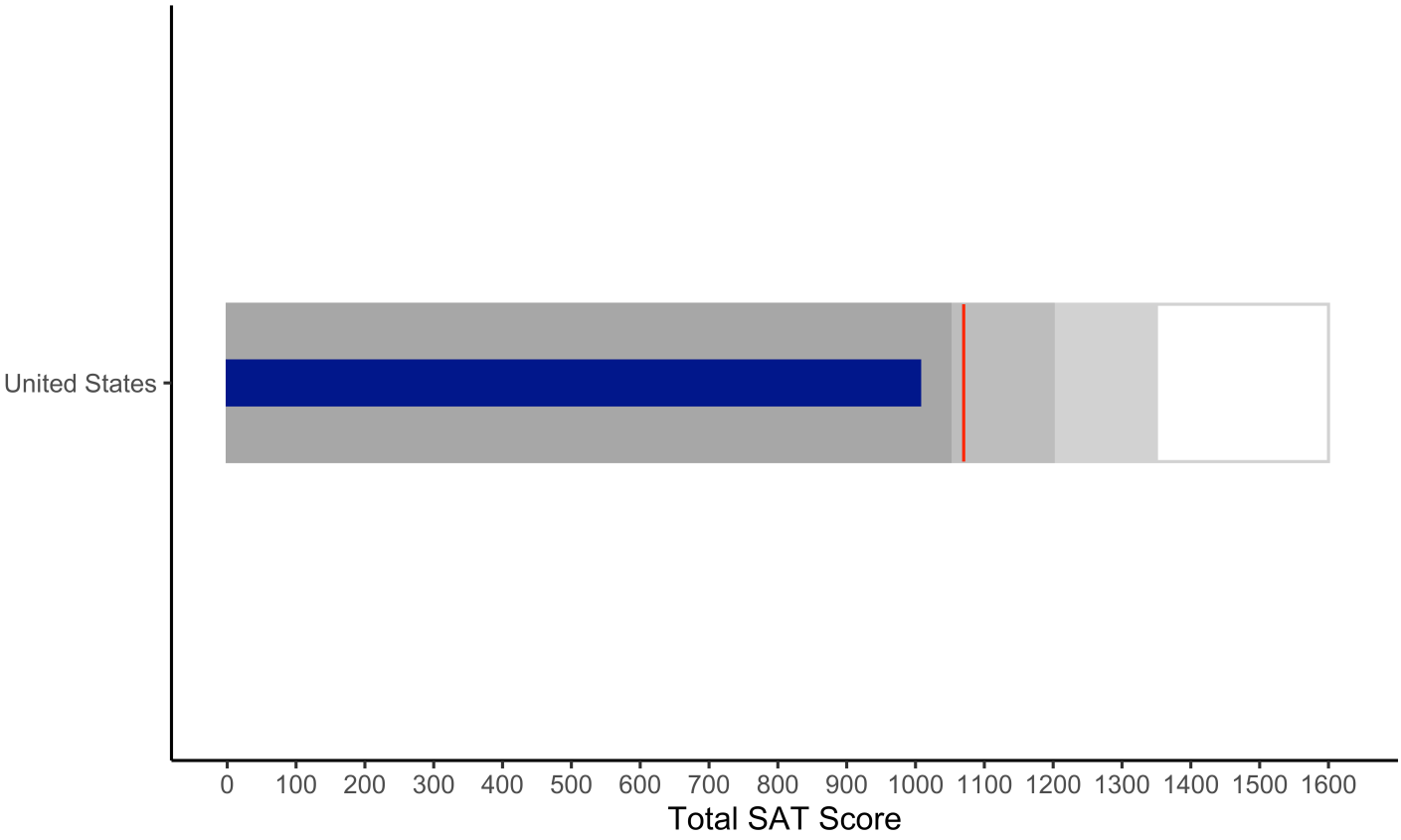
	state	reading	...	writing	percent_graduates_sat	pupil_staff_ratio	drop
	<fctr>	<int>	<int>	<int>	<int>	<dbl>	
1	United States	501	515	493	46	7.9	

1 row | 1-8 of 13 columns

```
# create boxplots
ggplot(data_2_sat) +
  geom_col(aes(x = state, y = 1600), color = 'light gray', fill = 'white', width = 0.
25) +
  geom_col(aes(x = state, y = sat_91), color = 'light gray', fill = 'light gray', wid
th = 0.25) +
  geom_col(aes(x = state, y = sat_74), color = 'gray', fill = 'gray', width = 0.25) +
  geom_col(aes(x = state, y = sat_49), color = 'dark gray', fill = 'dark gray', width
= 0.25) +
  geom_col(aes(x = state, y = Total_SAT_Score), color = 'dark blue', fill = 'dark blu
e', width = 0.07) +
  geom_segment(aes(x = 0.875, y = 1070), xend = 1.125, yend = 1070, color = 'red', )
+
  coord_flip() +
  ggtitle("United States SAT Scores") +
  labs(caption = "Source: Data Collected by Nathan Yau from National Center for Educa
tion Statistics",
       subtitle = "The target value is 1070 or thhe 50th percentile.  We can see the a
verage\nfor the United States falls below target",
       x = "",
       y = "Total SAT Score") +
  theme_classic() +
  theme(plot.title = element_text(face = "bold", size = 18),
        plot.subtitle = element_text(color = "light gray"),
        plot.caption = element_text(color = "light gray")
        ) +
   scale_y_continuous(breaks = seq(0, 1600, 100), limits = c(0, 1620))
```

United States SAT Scores

The target value is 1070 or thhe 50th percentile. We can see the average for the United States falls below target



Source: Data Collected by Nathan Yau from National Center for Education Statistics