Week 11&12 - Exercise 6.2

Name: Madhukar Ayachit

Date: 21 Feb 2022

Class: DSC-640

Assignment: histograms, box plots, and bullet charts

In [2]:
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import math
from matplotlib.ticker import FuncFormatter
import plotly
import plotly.figure_factory as ff
from pandas.plotting import parallel_coordinates
import numpy as np

%matplotlib inline
```

In [3]:
```
## read the data from excel file into a dataframe

education = pd.read_csv('Data/ex6-2/education.csv')
crime = pd.read_csv('Data/ex6-2/crimeratesbystate-formatted.csv')
birthrate = pd.read_csv('Data/ex6-2/birth-rate.csv')

# remove whitespaces from crime dataset (sine we have already encountered it)
education = education.applymap(lambda x: x.strip() if type(x) is str else x)
crime = crime.applymap(lambda x: x.strip() if type(x) is str else x)
birthrate = birthrate.applymap(lambda x: x.strip() if type(x) is str else x)
```

In [4]:
```
birthrate
```

Out[4]:

	Country	1960	1961	1962	1963	1964	1965	1966	1967	1968	...	
0	Aruba	36.400	35.179	33.863	32.459	30.994	29.513	28.069	26.721	25.518	...	1
1	Afghanistan	52.201	52.206	52.208	52.204	52.192	52.168	52.130	52.076	52.006	...	5
2	Angola	54.432	54.394	54.317	54.199	54.040	53.836	53.585	53.296	52.984	...	4
3	Albania	40.886	40.312	39.604	38.792	37.913	37.008	36.112	35.245	34.421	...	1
4	Netherlands Antilles	32.321	30.987	29.618	28.229	26.849	25.518	24.280	23.173	22.230	...	1
...	...	...	...	...	...	...	...	...	...	...	...	
229	Samoa	48.202	47.788	47.226	46.491	45.591	44.558	43.447	42.331	41.270	...	3
230	Yemen, Rep.	54.501	54.516	54.563	54.645	54.761	54.914	55.100	55.310	55.530	...	4
231	South Africa	42.267	41.993	41.610	41.112	40.520	39.883	39.268	38.734	38.317	...	2
232	Zambia	48.112	48.323	48.533	48.734	48.915	49.061	49.156	49.195	49.175	...	4

	Country	1960	1961	1962	1963	1964	1965	1966	1967	1968	...	
233	Zimbabwe	48.178	48.179	48.148	48.079	47.977	47.852	47.724	47.614	47.536	...	3

234 rows × 50 columns

In [5]:
```
crime
```

Out[5]:

	state	murder	forcible_rape	robbery	aggravated_assault	burglary	larceny_theft	r
0	United States	5.6	31.7	140.7	291.1	726.7	2286.3	
1	Alabama	8.2	34.3	141.4	247.8	953.8	2650.0	
2	Alaska	4.8	81.1	80.9	465.1	622.5	2599.1	
3	Arizona	7.5	33.8	144.4	327.4	948.4	2965.2	
4	Arkansas	6.7	42.9	91.1	386.8	1084.6	2711.2	
5	California	6.9	26.0	176.1	317.3	693.3	1916.5	
6	Colorado	3.7	43.4	84.6	264.7	744.8	2735.2	
7	Connecticut	2.9	20.0	113.0	138.6	437.1	1824.1	
8	Delaware	4.4	44.7	154.8	428.2	688.9	2144.0	
9	District of Columbia	35.4	30.2	672.1	721.3	649.7	2694.9	
10	Florida	5.0	37.1	169.4	496.6	926.3	2658.3	
11	Georgia	6.2	23.6	154.8	264.3	931.0	2751.1	
12	Hawaii	1.9	26.9	78.5	147.8	767.9	3308.4	
13	Idaho	2.4	40.4	18.6	195.4	564.4	1931.7	
14	Illinois	6.0	33.7	181.7	330.2	606.9	2164.8	
15	Indiana	5.7	29.6	108.6	179.9	697.6	2412.0	
16	Iowa	1.3	27.9	38.9	223.3	606.4	2042.7	
17	Kansas	3.7	38.4	65.3	280.0	689.2	2758.1	
18	Kentucky	4.6	34.0	88.4	139.8	634.0	1685.8	
19	Louisiana	9.9	31.4	118.0	435.1	870.6	2494.5	
20	Maine	1.4	24.7	24.4	61.7	478.5	1832.6	
21	Maryland	9.9	22.6	256.7	413.8	641.4	2294.3	
22	Massachusetts	2.7	27.1	119.0	308.1	541.1	1527.4	
23	Michigan	6.1	51.3	131.8	362.9	696.8	1917.8	
24	Minnesota	2.2	44.0	92.0	158.7	578.9	2226.9	
25	Mississippi	7.3	39.3	82.3	149.4	919.7	2083.9	
26	Missouri	6.9	28.0	124.1	366.4	738.3	2746.2	
27	Montana	1.9	32.2	18.9	228.5	389.2	2543.0	

	state	murder	forcible_rape	robbery	aggravated_assault	burglary	larceny_theft	r
28	Nebraska	2.5	32.9	59.1	192.5	532.4	2574.3	
29	Nevada	8.5	42.1	194.7	361.5	972.4	2153.9	
30	New Hampshire	1.4	30.9	27.4	72.3	317.0	1377.3	
31	New Jersey	4.8	13.9	151.6	184.4	447.1	1568.4	
32	New Mexico	7.4	54.1	98.7	541.9	1093.9	2639.9	
33	New York	4.5	18.9	182.7	239.7	353.3	1569.6	
34	North Carolina	6.7	26.5	145.5	289.4	1201.1	2546.2	
35	North Dakota	1.1	24.2	7.4	65.5	311.9	1500.3	
36	Ohio	5.1	39.8	163.1	143.4	872.8	2429.0	
37	Oklahoma	5.3	41.7	91.0	370.5	1006.0	2644.2	
38	Oregon	2.2	34.8	68.1	181.8	758.6	3112.2	
39	Pennsylvania	6.1	28.9	154.6	235.0	451.6	1729.1	
40	Rhode Island	3.2	29.8	72.1	146.1	494.2	1816.0	
41	South Carolina	7.4	42.5	132.1	579.0	1000.9	2954.1	
42	South Dakota	2.3	46.7	18.6	108.1	324.4	1343.7	
43	Tennessee	7.2	36.4	167.3	541.9	1026.9	2828.1	
44	Texas	6.2	37.2	156.6	329.8	961.6	2961.7	
45	Utah	2.3	37.3	44.3	143.4	606.2	2918.8	
46	Vermont	1.3	23.3	11.7	83.5	491.8	1686.1	
47	Virginia	6.1	22.7	99.2	154.8	392.1	2035.0	
48	Washington	3.3	44.7	92.1	205.8	959.7	3149.5	
49	West Virginia	4.4	17.7	44.6	206.1	621.2	1794.0	
50	Wisconsin	3.5	20.6	82.2	135.2	440.8	1992.8	
51	Wyoming	2.7	24.0	15.3	188.1	476.3	2533.9	

In [4]:
```
birthrate_hist = pd.melt(birthrate, id_vars="Country", var_name="Year", value_na
birthrate_hist["BirthRate_int"] = birthrate_hist["BirthRate"].apply(lambda x: ma
birthrate_hist.head()
```

Out[4]:

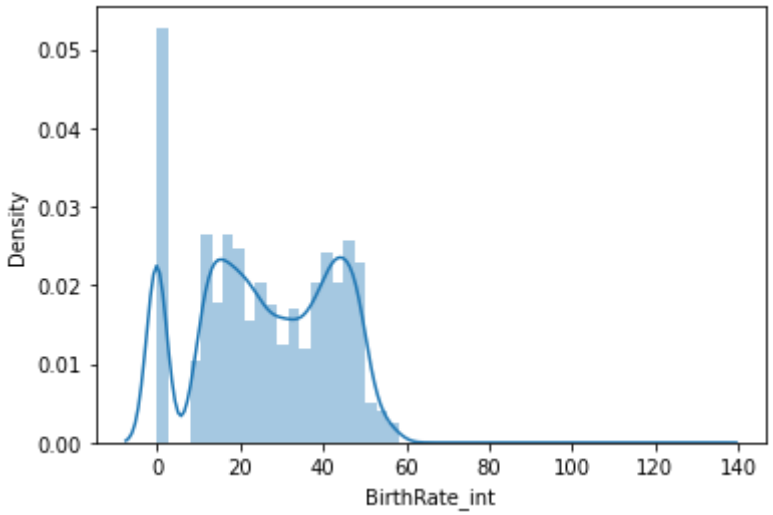
	Country	Year	BirthRate	BirthRate_int
0	Aruba	1960	36.400	37
1	Afghanistan	1960	52.201	53
2	Angola	1960	54.432	55
3	Albania	1960	40.886	41
4	Netherlands Antilles	1960	32.321	33

Histograms

In [5]:
```
sns.distplot( birthrate_hist["BirthRate_int"] )
```

```
/Users/madhukarayachit/opt/anaconda3/lib/python3.8/site-packages/seaborn/distrib
utions.py:2551: FutureWarning: `distplot` is a deprecated function and will be r
emoved in a future version. Please adapt your code to use either `displot` (a fi
gure-level function with similar flexibility) or `histplot` (an axes-level funct
ion for histograms).
  warnings.warn(msg, FutureWarning)
```

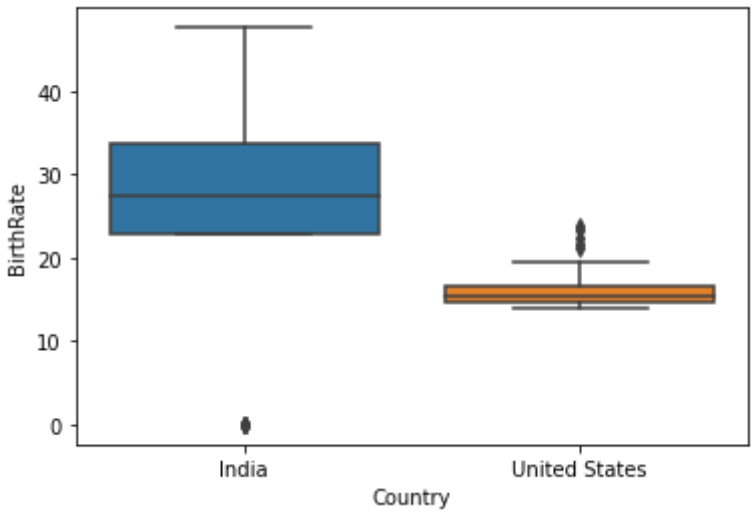
Out[5]:
```
<AxesSubplot:xlabel='BirthRate_int', ylabel='Density'>
```



Boxplot

In [6]:
```
birthrate_box = birthrate_hist[(birthrate_hist["Country"]=="United States") | (b
sns.boxplot(x = birthrate_box["Country"], y=birthrate_box["BirthRate"])
```

Out[6]:
```
<AxesSubplot:xlabel='Country', ylabel='BirthRate'>
```

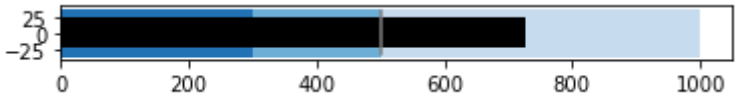


Bullet Charts

In [8]:
```
# transform data
```

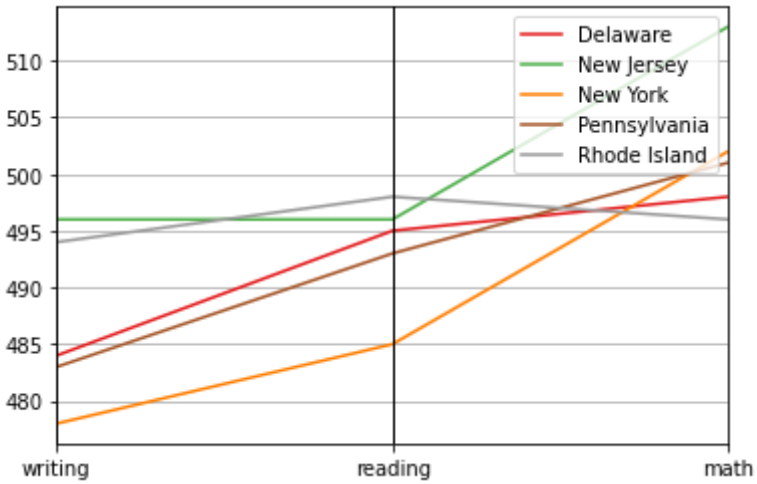
```
crime_bullet = crime[crime["state"]=="United States"][["state","burglary"]]
crime_bullet['target'] = 500
crime_bullet_tuple = [tuple(x) for x in crime_bullet.values][0]

# set parameter for bullet chart
limits = [300, 500, 1000]
palette = sns.color_palette("Blues_r", len(limits))
fig, ax = plt.subplots()
ax.set_aspect('equal')
#ax.set_yticks([1])
#ax.set_yticklabels(crime_bullet_tuple[0])

prev_limit = 0
for idx, lim in enumerate(limits):
    ax.barh([1], lim-prev_limit, left=prev_limit, height=75, color=palette[idx])
    prev_limit = lim

# draw the value we're measuring
ax.barh([1], crime_bullet_tuple[1], color='black', height=45)

ax.axvline(crime_bullet_tuple[2], color="gray", ymin=0.10, ymax=0.9)
```

Out[8]: <matplotlib.lines.Line2D at 0x7fa85633bfa0>



Parallel Coordinate plot

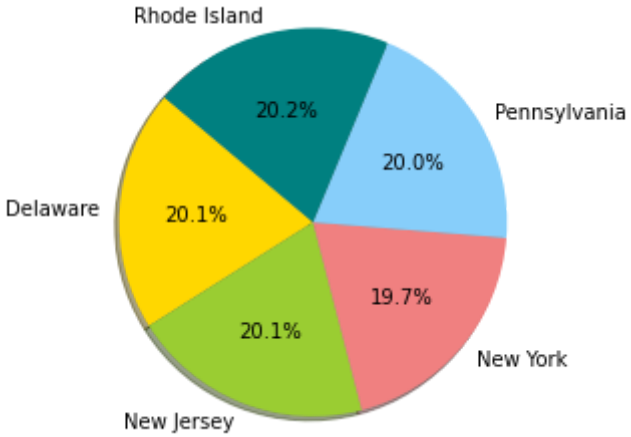
In [9]:
```
# transform data
education_parallel = education[education['state'].isin(['New York','New Jersey',

# make the plot
parallel_coordinates(education_parallel, 'state', colormap=plt.get_cmap("Set1"))
plt.show()
```



Pie Charts

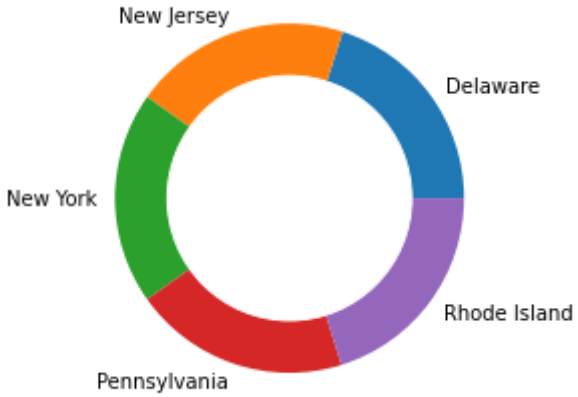
In [10]:
```
# transform data
education_pie = education_parallel[['state','reading']]
```

```
# set colors
colors = ['gold', 'yellowgreen', 'lightcoral', 'lightskyblue','teal']

# plot
plt.pie(education_pie['reading'], labels=education_pie['state'], colors=colors,
autopct='%1.1f%%', shadow=True, startangle=140)

plt.axis('equal')
plt.show()
```



Donut Charts

In [11]:
```
# transform data
education_donut = education_pie

# create a pieplot
plt.pie(education_donut['reading'], labels=education_donut['state'])

# add a circle at the center
my_circle=plt.Circle( (0,0), 0.7, color='white')
p=plt.gcf()
p.gca().add_artist(my_circle)

plt.show()
```



In [ ]:

histograms, box plots, and bullet charts

Code

Hide

```
# setting current working diirectory
setwd("/Users/madhukarayachit/DSC640")
```

Hide

```
#Load libraries
library('magrittr')
```

Hide

```


# load birth rate data
birthrate <- read.csv('Data/ex6-2/birth-rate.csv')

# load crime data
crime <- read.csv('Data/ex6-2/crimeratesbystate-formatted.csv')

# load education data
education <- read.csv('Data/ex6-2/education.csv')

# check column names
colnames(birthrate)
```

```
 [1] "Country" "X1960"   "X1961"   "X1962"   "X1963"   "X1964"   "X1965"   "X1966"   "X1
967"
[10] "X1968"   "X1969"   "X1970"   "X1971"   "X1972"   "X1973"   "X1974"   "X1975"   "X1
976"
[19] "X1977"   "X1978"   "X1979"   "X1980"   "X1981"   "X1982"   "X1983"   "X1984"   "X1
985"
[28] "X1986"   "X1987"   "X1988"   "X1989"   "X1990"   "X1991"   "X1992"   "X1993"   "X1
994"
[37] "X1995"   "X1996"   "X1997"   "X1998"   "X1999"   "X2000"   "X2001"   "X2002"   "X2
003"
[46] "X2004"   "X2005"   "X2006"   "X2007"   "X2008"
```

Hide

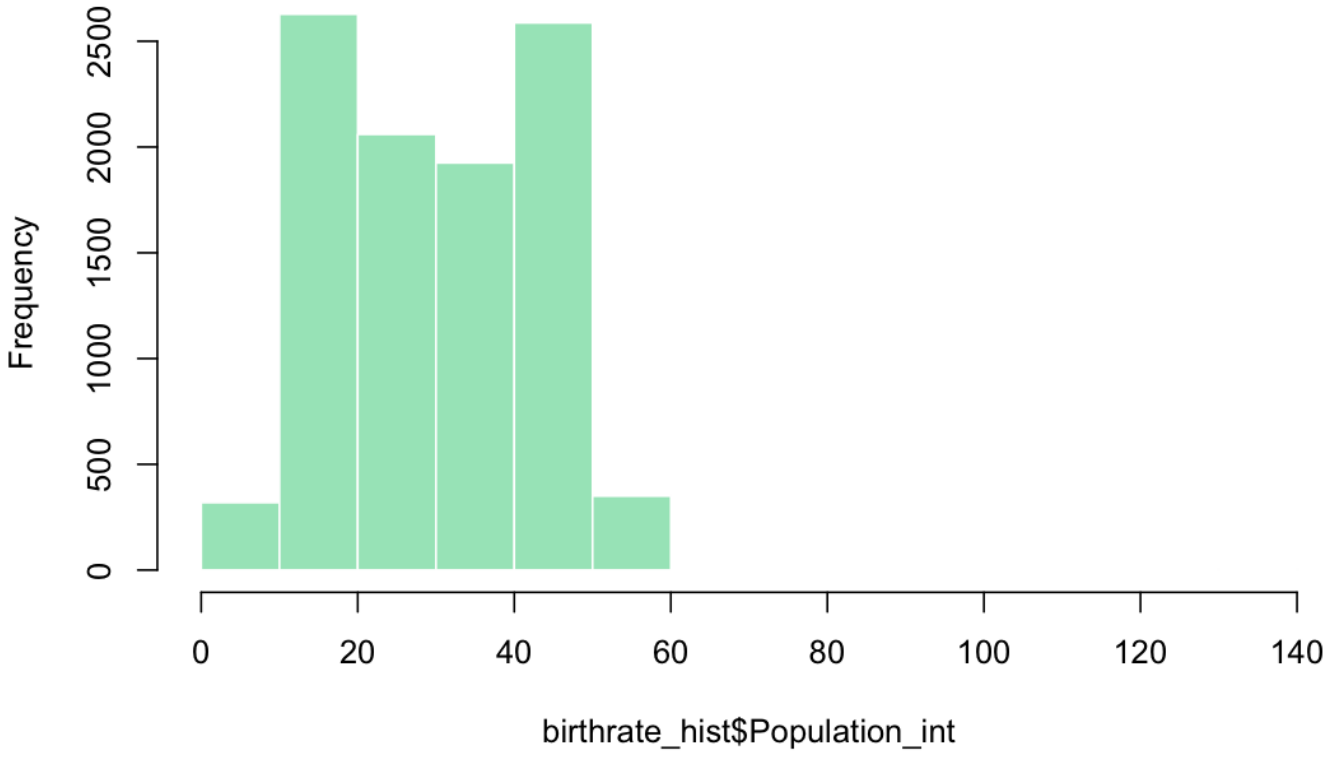
```
# format year columns
colnames(birthrate) <- gsub("X", "", colnames(birthrate))

# check column names
colnames(birthrate)
```

```
 [1] "Country" "1960"    "1961"    "1962"    "1963"    "1964"    "1965"    "1966"    "19
67"
[10] "1968"    "1969"    "1970"    "1971"    "1972"    "1973"    "1974"    "1975"    "19
76"
[19] "1977"    "1978"    "1979"    "1980"    "1981"    "1982"    "1983"    "1984"    "19
85"
[28] "1986"    "1987"    "1988"    "1989"    "1990"    "1991"    "1992"    "1993"    "19
94"
[37] "1995"    "1996"    "1997"    "1998"    "1999"    "2000"    "2001"    "2002"    "20
03"
[46] "2004"    "2005"    "2006"    "2007"    "2008"
```

Histogram

Hide

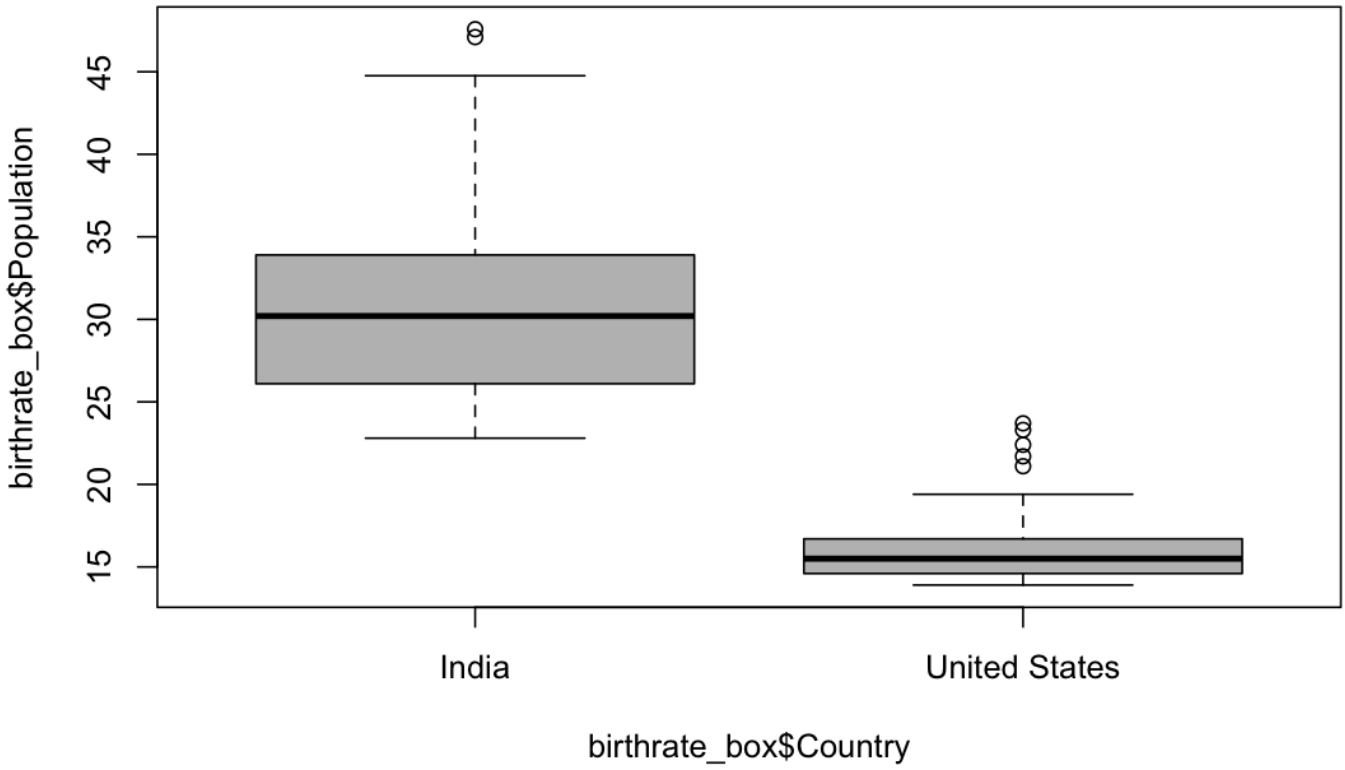
```
options(repr.plot.width = 4, repr.plot.height = 4)

# create pivotted data for plotting
birthrate_hist <- reshape2::melt(birthrate, id=c("Country")) %>%
  dplyr::mutate("Country" = as.character(Country),
                "Year" = as.character(variable),
                "Population" = value,
                "Population_int" = ceiling(value)) %>%
  dplyr::select(c("Country","Year","Population", "Population_int"))

# create histogram of population data
hist(birthrate_hist$Population_int, col=rgb(0.2,0.8,0.5,0.5) , border=F , main="")
```



Box plot

Hide

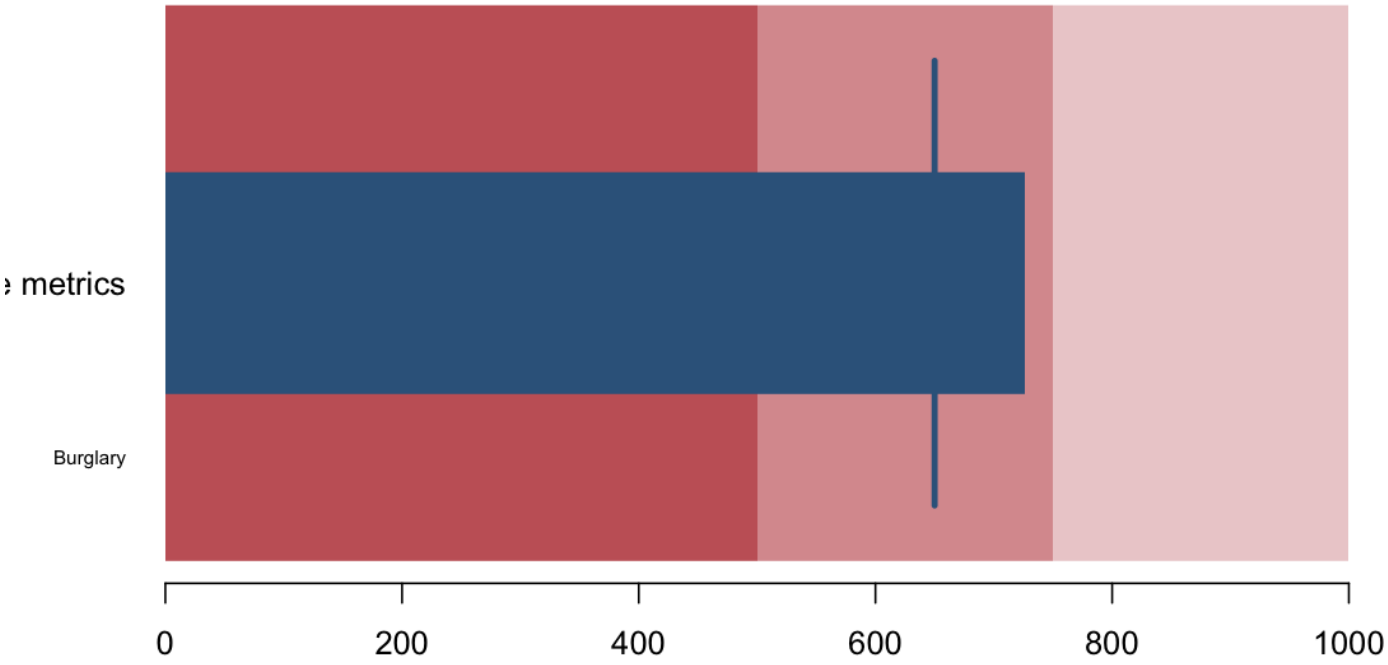
```
# create box plot of population data
birthrate_box <- birthrate_hist %>%
  dplyr::filter(Country %in% c("United States", "India"))

boxplot(birthrate_box$Population ~ birthrate_box$Country , col="grey")
```



Bullet chart

Hide

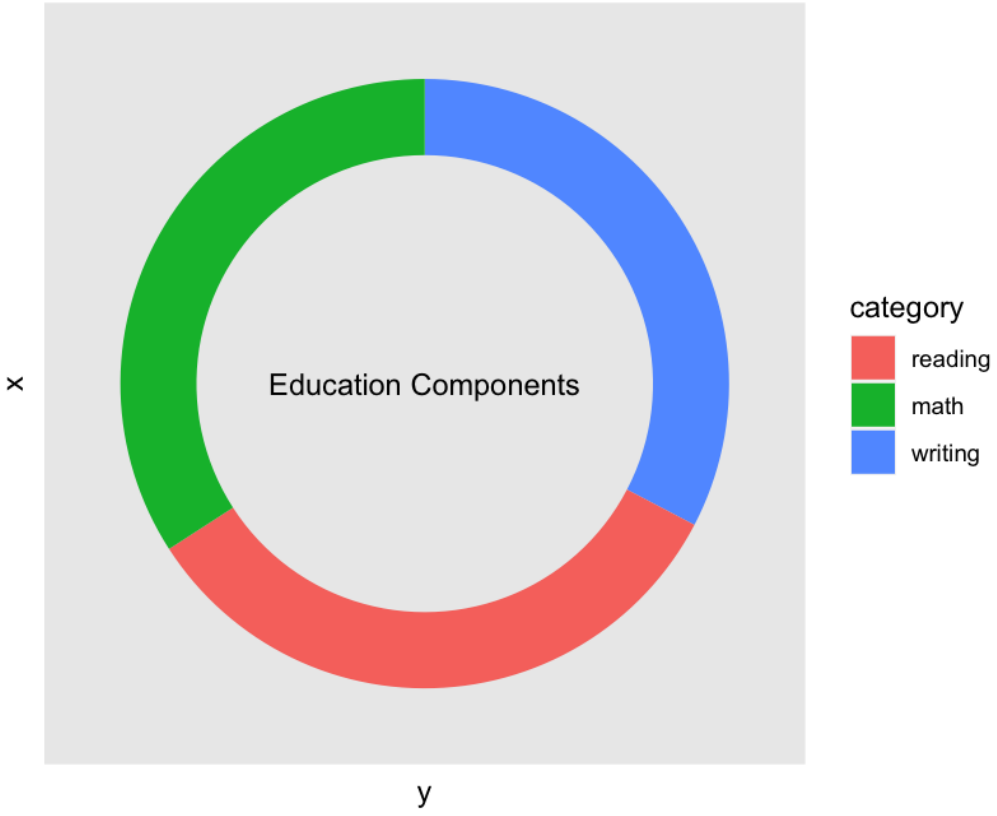
```
source("BulletGraph.r", local=TRUE)

# create bullet chart with crime data
crime_bullet <- crime %>%
  dplyr::filter(stringr::str_trim(state, "both") == "United States") %>%
  dplyr::select(c(state, burglary))

bulletgraph(x=crime_bullet$burglary,ref=650,limits=c(0,500,750,1000),
            name= "USA Crime metrics",subname="Burglary",
            col="steelblue4",shades="firebrick")
```



Donut Chart

Hide

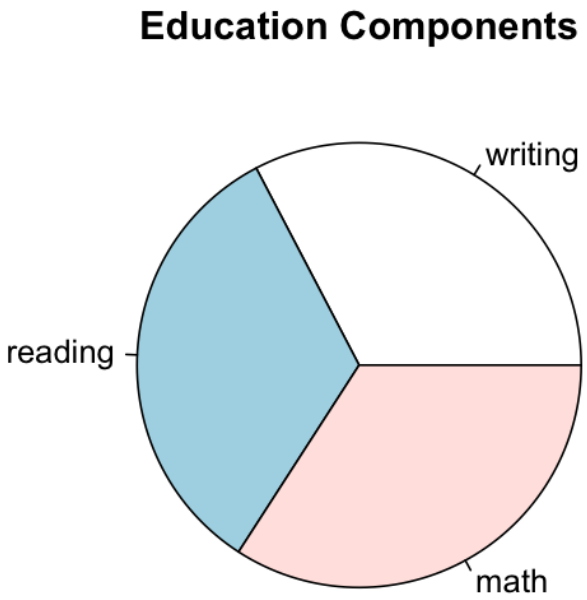
```
# donut chart using USA crime data
education_donut <- education %>%
  dplyr::filter(stringr::str_trim(state, "both") == "United States") %>%
  reshape2::melt(id=c("state")) %>%
  dplyr::rename("category" = variable) %>%
  dplyr::filter(category %in% c("reading","math","writing")) %>%
  dplyr::select(-state)

# add addition columns, needed for drawing with geom_rect
education_donut$fraction = education_donut$value / sum(education_donut$value)
education_donut = education_donut[order(education_donut$fraction), ]
education_donut$ymax = cumsum(education_donut$fraction)
education_donut$ymin = c(0, head(education_donut$ymax, n=-1))

# make the plot
ggplot2::ggplot(education_donut, ggplot2::aes(fill=category, ymax=ymax, ymin=ymin, xmax=
4, xmin=3)) +
  ggplot2::geom_rect() +
  ggplot2::coord_polar(theta="y") +
  ggplot2::xlim(c(0, 4)) +
  ggplot2::theme(panel.grid=ggplot2::element_blank()) +
  ggplot2::theme(axis.text=ggplot2::element_blank()) +
  ggplot2::theme(axis.ticks=ggplot2::element_blank()) +
  ggplot2::annotate("text", x = 0, y = 0, label = "Education Components") +
  ggplot2::labs(title="")
```



Pie Chart

Hide

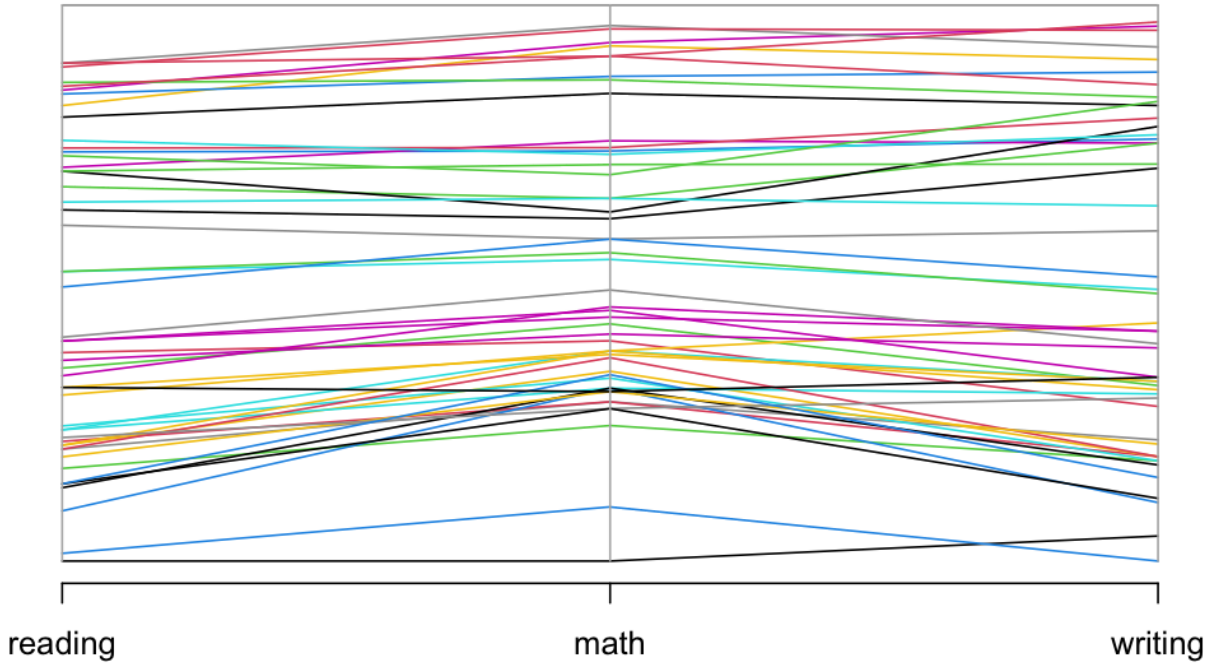
```
# pie chart
slices <- education_donut$value
lbls <- education_donut$category
pie(slices, labels = lbls, main="Education Components")
```



Parallel Plot

Hide

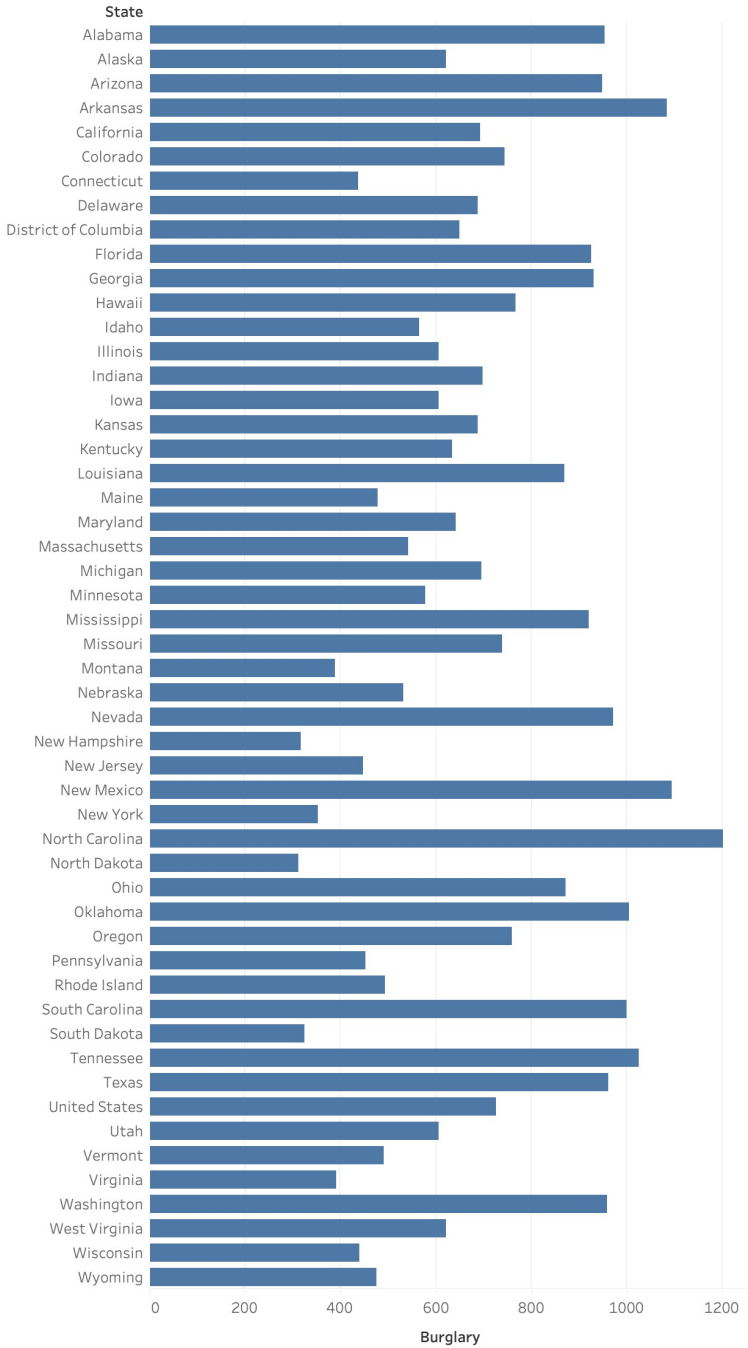
```
# parallel plot
education_parallel <- education %>%
  dplyr::filter(stringr::str_trim(state, "both") != "United States")

# vector color
my_colors <- as.numeric(factor(c(education_parallel$state)))


# make the graph
MASS::parcoord(education_parallel[,c(2:4)] , col= my_colors  )
```


reading
math
writing

MA_W_11_12

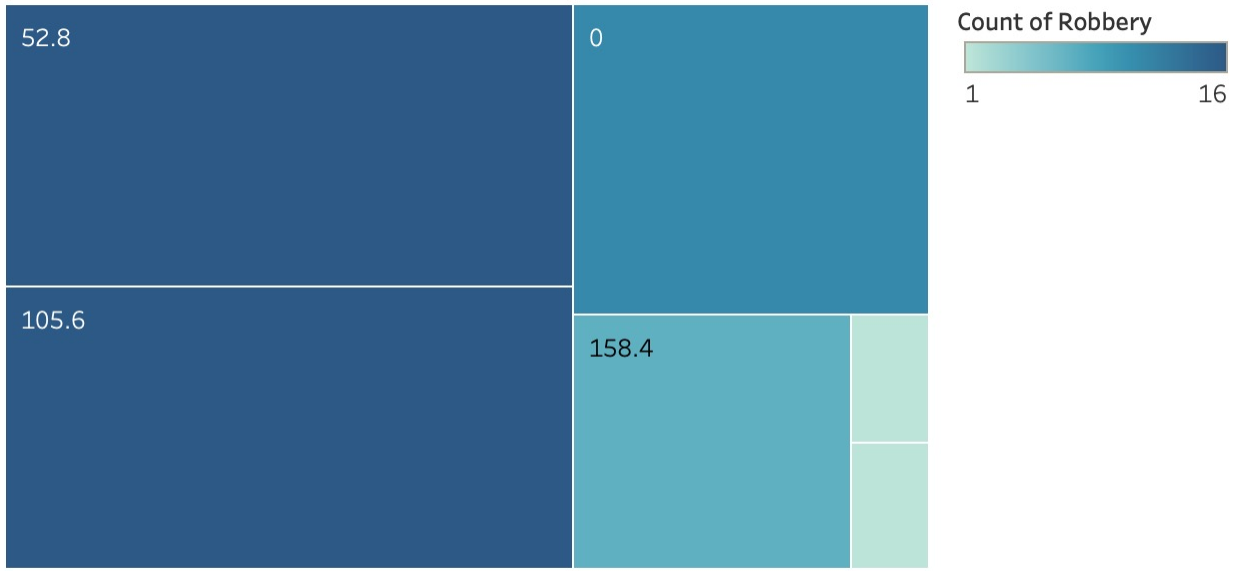
File created on: 2/27/22 8:12:01 PM EST

Horizontal Bars : Burglary



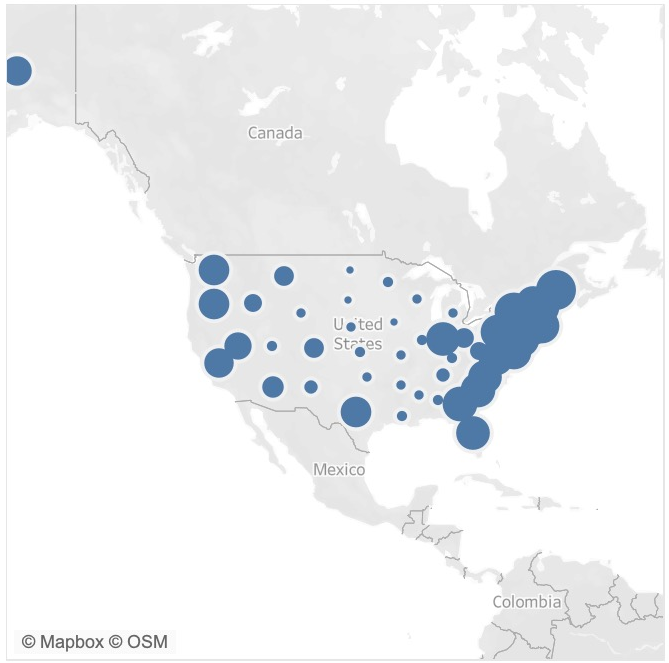
Sum of Burglary for each State.

TreeMap : Robery



Robbery (bin). Color shows count of Robbery. Size shows count of Robbery. The marks are labeled by Robbery (bin).

Maps : Grauduation

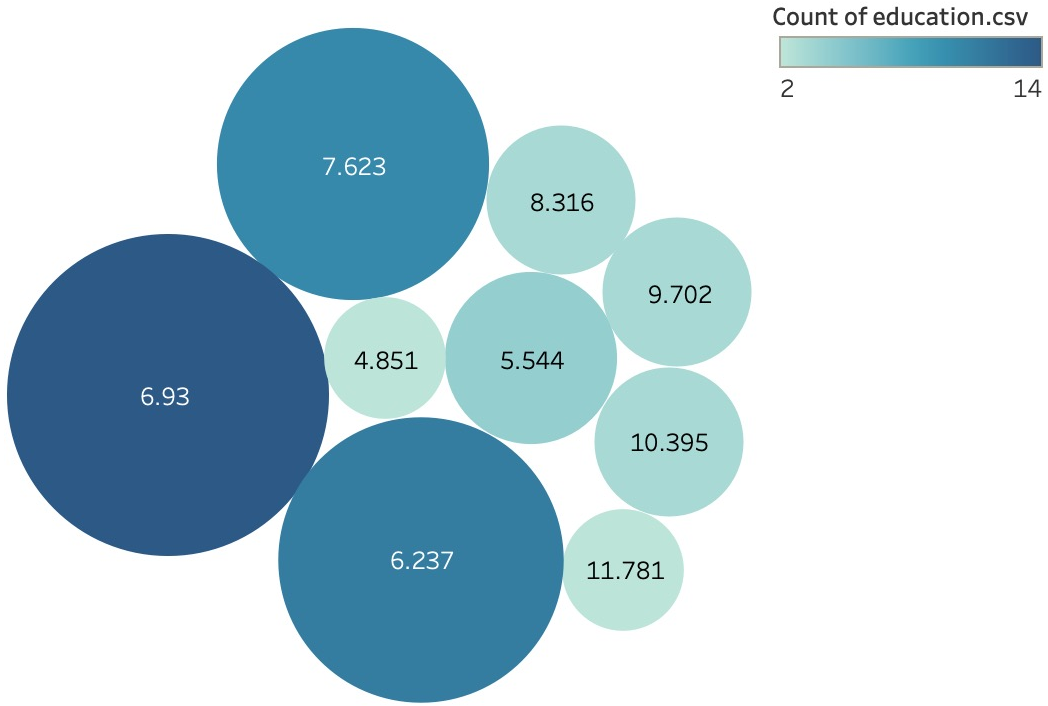


Percent Graduates Sat

- 3
- 20
- 40
- 60
- 80
- 90

Map based on Longitude (generated) and Latitude (generated). Size shows sum of Percent Graduates Sat. Details are shown for State (Education.Csv).

Bubble : Pupil Staff Ratio



Pupil Staff Ratio (bin). Color shows count of education.csv. Size shows count of Pupil Staff Ratio. The marks are labeled by Pupil Staff Ratio (bin).