DSC640-T301 Data Presentation & Visualization (2247-1)

Assignment: Week 9 & 10 Exercise

Author: Zemelak Goraga;

Date: 8/9/2024

About the 6 Datasets

Birth Rate Dataset: This dataset contains the birth rates for various countries from 1960 to 2008. It provides annual data, allowing for analysis of trends and patterns in birth rates across different regions over time.

Crime Rates by State (crimeratesbystate-formatted): This dataset includes crime statistics for different states in the United States. It covers categories such as murder, forcible rape, robbery, aggravated assault, burglary, larceny-theft, and motor vehicle theft, offering insights into crime rates and their variations across states.

Education Dataset: The education dataset provides information on academic performance and school metrics for different states. It includes data on reading, math, and writing scores, percentage of graduates taking the SAT, pupil-staff ratio, and dropout rates, allowing for comparative analysis of educational outcomes.

Staff, Enrollment, and Pupil/Staff Ratios (tabn084): This dataset presents data on public elementary and secondary school systems, including staff counts, student enrollment, and pupil/staff ratios across different states from 2000 to 2007. It helps understand staffing trends and resource allocation in education.

Public High School Graduates and Dropouts (tabn106): This dataset provides data on high school graduates and dropout rates by race/ethnicity and state for the 2006-07 academic year. It includes total counts and percentages, offering insights into demographic disparities in education.

SAT Mean Scores (tabn146): This dataset contains SAT scores of college-bound seniors by state, covering critical reading, mathematics, and writing sections. It also includes the percentage of graduates taking the SAT, enabling analysis of student performance and participation in standardized testing.

Research Questions

1. How do crime rates correlate with educational outcomes across different states?
2. What trends can be observed in birth rates across countries over the years, and how do these trends relate to educational and economic indicators?
3. What are the variations in high school graduation rates and dropout rates across different racial/ethnic groups, and how do they relate to standardized test scores (SAT)?

In [11]:
```
# Import Necessary Libraries

import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import plotly.express as px
from plotly.subplots import make_subplots
import plotly.graph_objects as go
import warnings
warnings.filterwarnings('ignore')
```

In [ ]:
```
pip install xlrd==2.0.1
```

In [5]:
```
# Import datasets
df_birth_rate = pd.read_csv('birth-rate.csv')
df_crime_rates = pd.read_csv('crimeratesbystate-formatted.csv')
df_education = pd.read_csv('education.csv')
df_tabn084 = pd.read_excel('tabn084.xls')
df_tabn106 = pd.read_excel('tabn106.xls')
df_tabn146 = pd.read_excel('tabn146.xls')
```

In [ ]:
```
# Inspect datasets
print(df_birth_rate.head())
print(df_crime_rates.head())
print(df_education.head())
print(df_tabn084.head())
print(df_tabn106.head())
print(df_tabn146.head())
```

In [12]:
```
# Data wrangling steps (e.g., handling missing values, data type conversions, etc.)
# Example: Fill missing values, if any
df_birth_rate.fillna(method='ffill', inplace=True)
df_crime_rates.fillna(method='ffill', inplace=True)
df_education.fillna(method='ffill', inplace=True)
df_tabn084.fillna(method='ffill', inplace=True)
df_tabn106.fillna(method='ffill', inplace=True)
df_tabn146.fillna(method='ffill', inplace=True)
```

In [19]:
```
# Print column names to check for the correct column name
print(df_birth_rate.columns)

# Check if 'Year' column exists
if 'Year' in df_birth_rate.columns:
```

```
    df_birth_rate['Year'] = df_birth_rate['Year'].astype(int)
else:
    print("The 'Year' column does not exist in the DataFrame.")
```

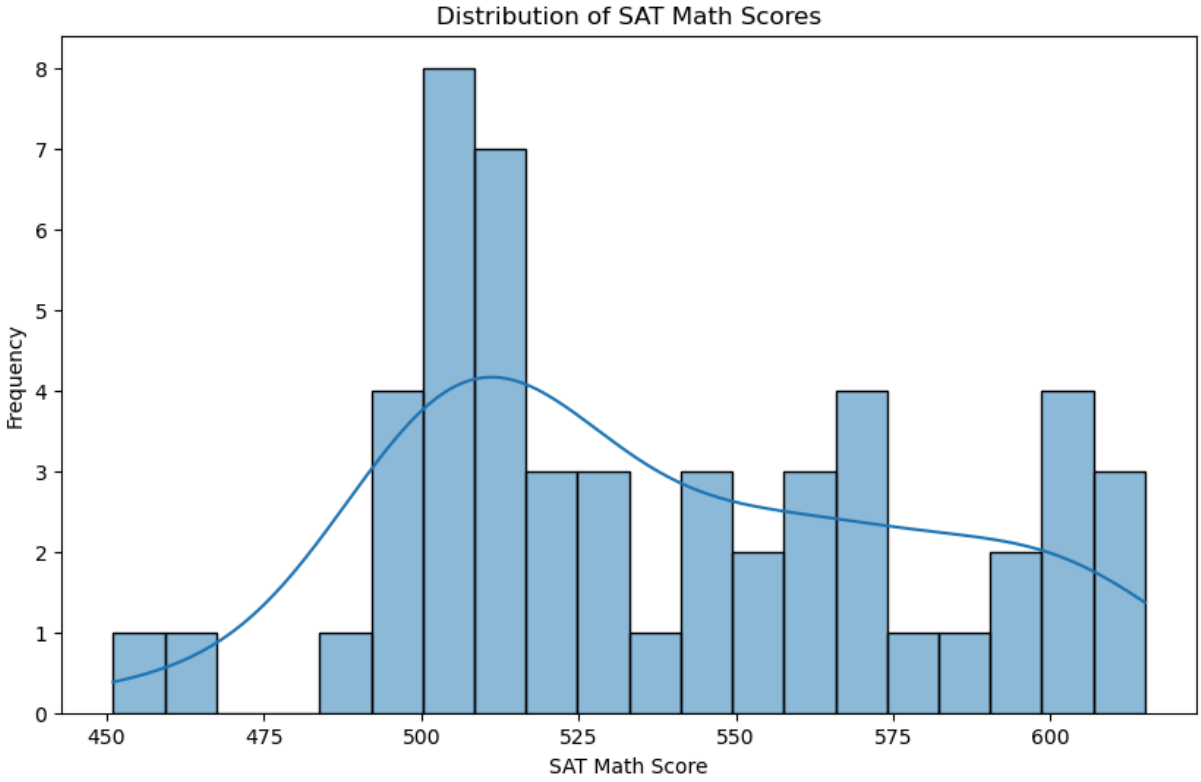
```
Index(['Country', '1960', '1961', '1962', '1963', '1964', '1965', '1966',
       '1967', '1968', '1969', '1970', '1971', '1972', '1973', '1974', '1975',
       '1976', '1977', '1978', '1979', '1980', '1981', '1982', '1983', '1984',
       '1985', '1986', '1987', '1988', '1989', '1990', '1991', '1992', '1993',
       '1994', '1995', '1996', '1997', '1998', '1999', '2000', '2001', '2002',
       '2003', '2004', '2005', '2006', '2007', '2008'],
      dtype='object')
The 'Year' column does not exist in the DataFrame.
```

In [20]:
```
# Reshape the DataFrame using melt
df_birth_rate_melted = df_birth_rate.melt(id_vars=['Country'],
                                          var_name='Year',
                                          value_name='BirthRate')

# Convert the 'Year' column to integers
df_birth_rate_melted['Year'] = df_birth_rate_melted['Year'].astype(int)

# Display the first few rows of the reshaped DataFrame
print(df_birth_rate_melted.head())
```

```
                Country  Year  BirthRate
0                 Aruba  1960     36.400
1           Afghanistan  1960     52.201
2                Angola  1960     54.432
3               Albania  1960     40.886
4  Netherlands Antilles  1960     32.321
```

Python Visualizations

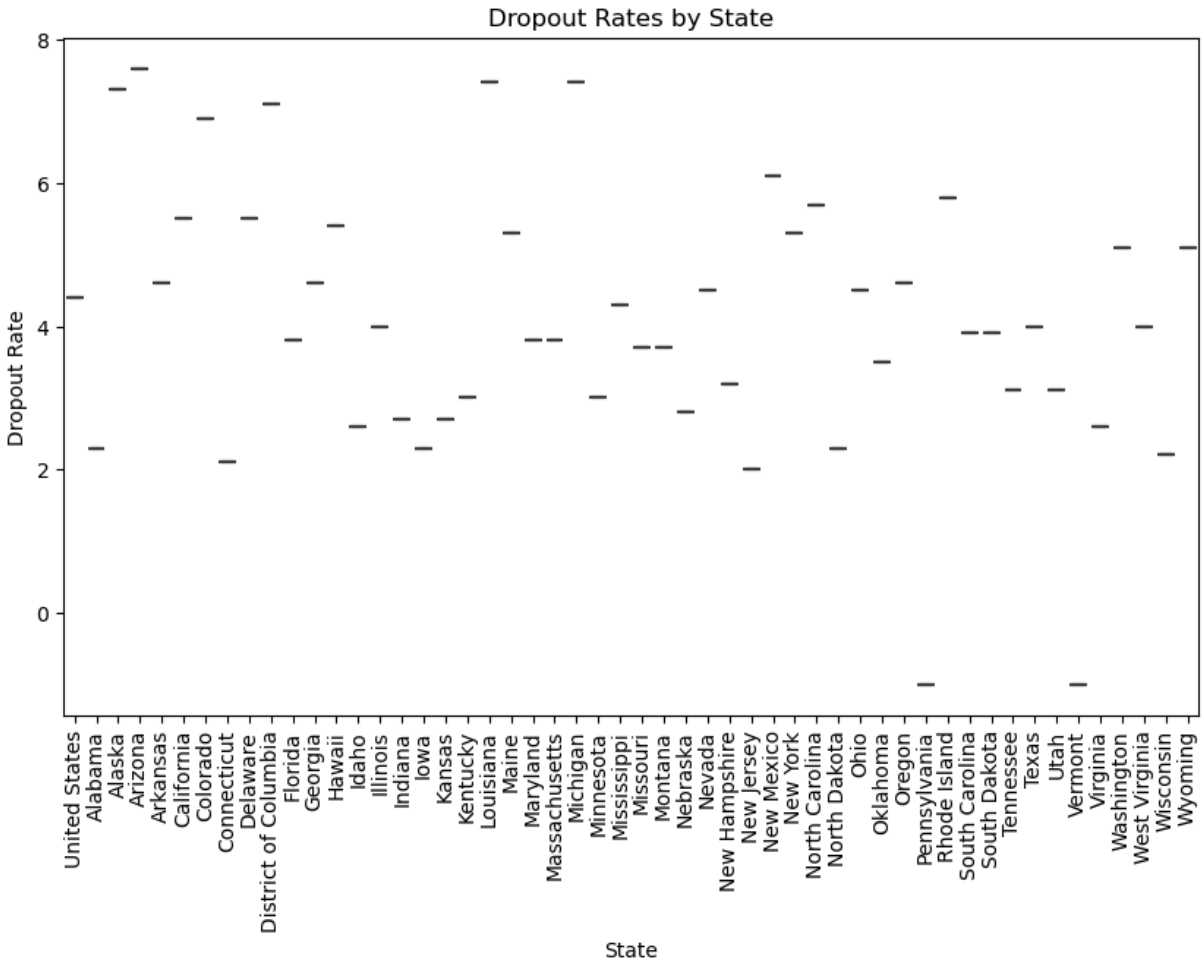
In [14]:
```
# Histogram (Python)

import matplotlib.pyplot as plt
import seaborn as sns

# Assuming df_education is the dataset used
plt.figure(figsize=(10, 6))
sns.histplot(df_education['math'], bins=20, kde=True)
plt.title('Distribution of SAT Math Scores')
plt.xlabel('SAT Math Score')
plt.ylabel('Frequency')
plt.show()
```



Purpose

A histogram displays the distribution of a dataset by grouping data into bins and showing the frequency of each bin. It is essential for understanding the spread and central tendency of a dataset, such as the distribution of SAT scores across states. Relative Importance: Histograms are vital for identifying patterns, such as the concentration of scores in specific ranges, and for detecting outliers.

In [15]:
```
# Box Plot (Python)

plt.figure(figsize=(10, 6))
sns.boxplot(x='state', y='dropout_rate', data=df_education)
plt.title('Dropout Rates by State')
plt.xlabel('State')
plt.ylabel('Dropout Rate')
plt.xticks(rotation=90)
plt.show()
```

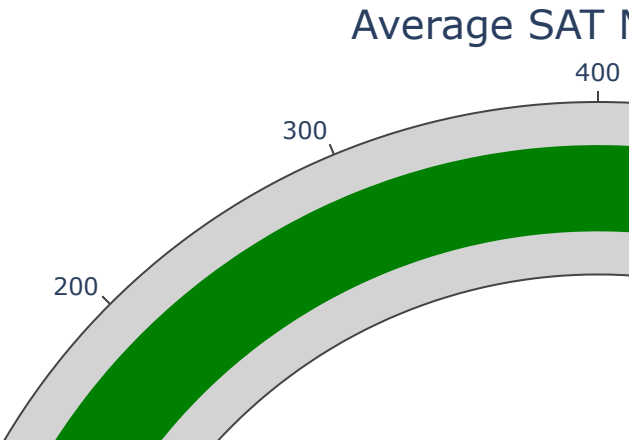


Purpose

A box plot visualizes the distribution of a dataset through its quartiles, highlighting the median, interquartile range, and potential outliers. It is particularly useful for comparing distributions across multiple groups, such as dropout rates across different states. Relative Importance: Box plots provide a clear summary of the central tendency and variability, making them valuable for comparative analysis.

In [16]:
```python
# Bullet Chart (Python)

import plotly.graph_objects as go

# Example data
target = 500
actual = df_education['math'].mean()

fig = go.Figure(go.Indicator(
    mode="gauge+number+delta",
    value=actual,
    title={'text': "Average SAT Math Score"},
    gauge={'axis': {'range': [0, 800]},
           'steps': [{'range': [0, 500], 'color': "lightgray"},
```

```
                    {'range': [500, 800], 'color': "gray"}],
           'threshold': {'line': {'color': "red", 'width': 4}, 'thickness': 0.75, '
fig.show()
```

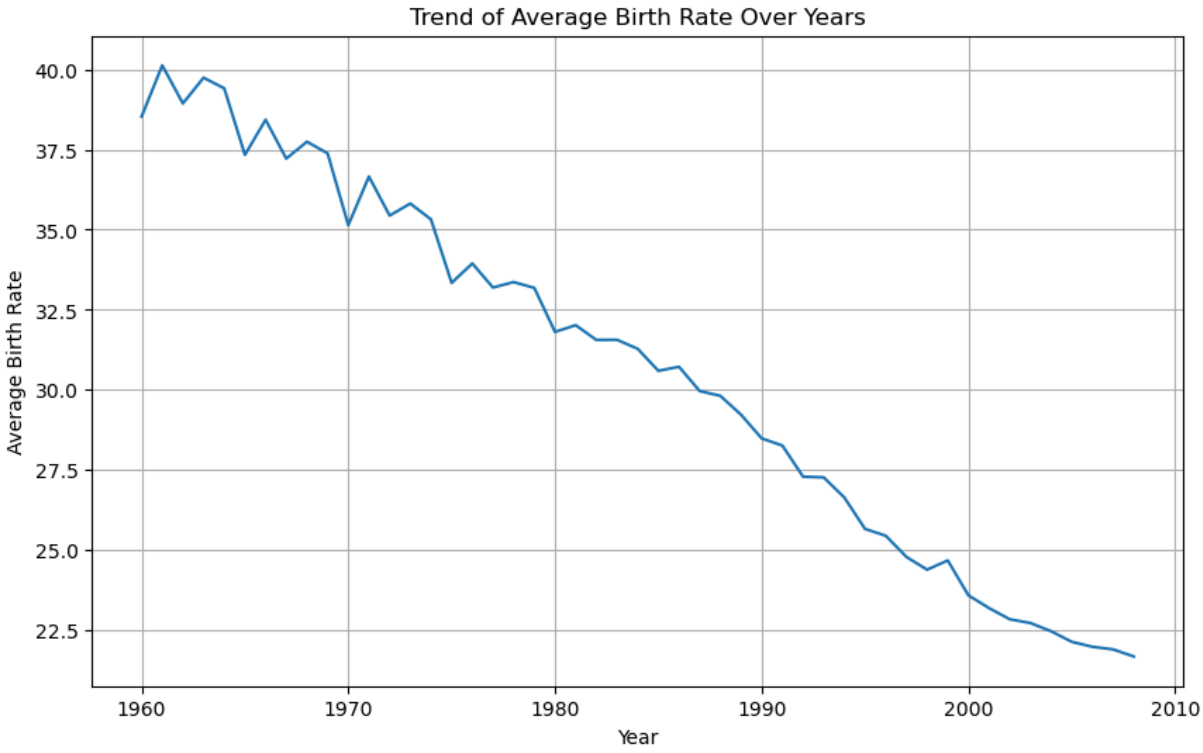


Purpose

A bullet chart compares a primary measure against one or more other measures, such as a target or historical performance. It can be used to display the actual versus target SAT scores. Relative Importance: Bullet charts are crucial for performance monitoring, as they provide a straightforward way to assess whether goals are being met.

In [24]:
```
import matplotlib.pyplot as plt

# Group by 'Year' and calculate the mean birth rate for each year
average_birth_rate_per_year = df_birth_rate_melted.groupby('Year')['BirthRate'].mea

# Plotting
plt.figure(figsize=(10, 6))
average_birth_rate_per_year.plot()
plt.title('Trend of Average Birth Rate Over Years')
plt.xlabel('Year')
```

```
plt.ylabel('Average Birth Rate')
plt.grid(True)
plt.show()
```



Purpose

A line chart displays data points connected by straight lines, representing trends over time. It is ideal for visualizing changes in birth rates over the years.

Relative Importance: Line charts are essential for identifying trends and patterns in time-series data, providing insights into long-term changes and cyclical patterns.

R Visualizations

DSC640-Week9&10_excercise_Visualization

Zemelak Goraga

2024-08-04

```
# Install necessary packages
if (!requireNamespace("sf", quietly = TRUE)) install.packages("sf")
if (!requireNamespace("ggplot2", quietly = TRUE)) install.packages("ggplot2")
if (!requireNamespace("viridis", quietly = TRUE)) install.packages("viridis")
if (!requireNamespace("utils", quietly = TRUE)) install.packages("utils")
```

```
# Load necessary libraries
library(sf)
library(ggplot2)
library(viridis)
library(utils)
```

```
# Set working directory
setwd("C:/Users/zemel/Burn/")
```

```
# Check current working directory
getwd()
```

```
## [1] "C:/Users/zemel/Burn"
```

```
# Set working directory (ensure this is correct for your system)
setwd("C:/Users/zemel/Burn/")
```

```
# Import datasets
df_birth_rate <- read_csv('birth-rate.csv')
```

```
## Rows: 234 Columns: 50
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (1): Country
## dbl (49): 1960, 1961, 1962, 1963, 1964, 1965, 1966, 1967, 1968, 1969, 1970, ...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
df_crime_rates <- read_csv('crimeratesbystate-formatted.csv')
```

```
## Rows: 52 Columns: 8
## -- Column specification --------------------------------------------------------
```

```
## Delimiter: ","
## chr (1): state
## dbl (7): murder, forcible_rape, robbery, aggravated_assault, burglary, larce...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
df_education <- read_csv('education.csv')
```

```
## Rows: 52 Columns: 7
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): state
## dbl (6): reading, math, writing, percent_graduates_sat, pupil_staff_ratio, d...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
df_tabn084 <- read_excel('tabn084.xls')
```

```
## New names:
## * `` -> `...2`
## * `` -> `...3`
## * `` -> `...4`
## * `` -> `...5`
## * `` -> `...6`
## * `` -> `...7`
## * `` -> `...8`
## * `` -> `...9`
## * `` -> `...10`
## * `` -> `...11`
## * `` -> `...12`
## * `` -> `...13`
## * `` -> `...14`
## * `` -> `...15`
## * `` -> `...16`
## * `` -> `...17`
## * `` -> `...18`
## * `` -> `...19`
## * `` -> `...20`
## * `` -> `...21`
## * `` -> `...22`
## * `` -> `...23`
## * `` -> `...24`
## * `` -> `...25`
## * `` -> `...26`
## * `` -> `...27`
## * `` -> `...28`
```

```
df_tabn106 <- read_excel('tabn106.xls')
```

```
## New names:
```

```
## * `` -> `...2`
## * `` -> `...3`
## * `` -> `...4`
## * `` -> `...5`
## * `` -> `...6`
## * `` -> `...7`
## * `` -> `...8`
## * `` -> `...9`
## * `` -> `...10`
## * `` -> `...11`
## * `` -> `...12`
## * `` -> `...13`
## * `` -> `...14`
```

```
df_tabn146 <- read_excel('tabn146.xls')
```

```
## New names:
## * `` -> `...2`
## * `` -> `...3`
## * `` -> `...4`
## * `` -> `...5`
## * `` -> `...6`
## * `` -> `...7`
## * `` -> `...8`
## * `` -> `...9`
## * `` -> `...10`
## * `` -> `...11`
## * `` -> `...12`
## * `` -> `...13`
## * `` -> `...14`
## * `` -> `...15`
## * `` -> `...16`
## * `` -> `...17`
## * `` -> `...18`
```

```
# Inspect datasets
print(head(df_birth_rate))
```

```
## # A tibble: 6 x 50
##   Country  `1960` `1961` `1962` `1963` `1964` `1965` `1966` `1967` `1968` `1969`
##   <chr>     <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>
## 1 Aruba      36.4   35.2   33.9   32.5   31.0   29.5   28.1   26.7   25.5   24.5
## 2 Afghani~   52.2   52.2   52.2   52.2   52.2   52.2   52.1   52.1   52.0   51.9
## 3 Angola     54.4   54.4   54.3   54.2   54.0   53.8   53.6   53.3   53.0   52.7
## 4 Albania    40.9   40.3   39.6   38.8   37.9   37.0   36.1   35.2   34.4   33.7
## 5 Netherl~   32.3   31.0   29.6   28.2   26.8   25.5   24.3   23.2   22.2   21.5
## 6 Arab Wo~   47.6   NA     NA     NA     NA     46.6   NA     NA     NA     NA
## # i 39 more variables: `1970` <dbl>, `1971` <dbl>, `1972` <dbl>, `1973` <dbl>,
## #   `1974` <dbl>, `1975` <dbl>, `1976` <dbl>, `1977` <dbl>, `1978` <dbl>,
## #   `1979` <dbl>, `1980` <dbl>, `1981` <dbl>, `1982` <dbl>, `1983` <dbl>,
## #   `1984` <dbl>, `1985` <dbl>, `1986` <dbl>, `1987` <dbl>, `1988` <dbl>,
## #   `1989` <dbl>, `1990` <dbl>, `1991` <dbl>, `1992` <dbl>, `1993` <dbl>,
## #   `1994` <dbl>, `1995` <dbl>, `1996` <dbl>, `1997` <dbl>, `1998` <dbl>,
## #   `1999` <dbl>, `2000` <dbl>, `2001` <dbl>, `2002` <dbl>, `2003` <dbl>, ...
```

```
print(head(df_crime_rates))
```

```
## # A tibble: 6 x 8
##   state   murder forcible_rape robbery aggravated_assault burglary larceny_theft
##   <chr>    <dbl>         <dbl>   <dbl>              <dbl>    <dbl>         <dbl>
## 1 United~    5.6          31.7   141.                291.     727.         2286.
## 2 Alabama    8.2          34.3   141.                248.     954.         2650
## 3 Alaska     4.8          81.1    80.9               465.     622.         2599.
## 4 Arizona    7.5          33.8   144.                327.     948.         2965.
## 5 Arkans~    6.7          42.9    91.1               387.    1085.         2711.
## 6 Califo~    6.9          26     176.                317.     693.         1916.
## # i 1 more variable: motor_vehicle_theft <dbl>
```

```
print(head(df_education))
```

```
## # A tibble: 6 x 7
##   state         reading  math writing percent_graduates_sat pupil_staff_ratio
##   <chr>           <dbl> <dbl>   <dbl>                 <dbl>             <dbl>
## 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
## # i 1 more variable: dropout_rate <dbl>
```

```
print(head(df_tabn084))
```

```
## # A tibble: 6 x 28
##   Table 84. Staff, enrol~1 ...2  ...3  ...4  ...5  ...6  ...7  ...8  ...9  ...10
##   <chr>                    <chr> <chr> <chr> <chr> <chr> <chr> <chr> <chr> <chr>
## 1 "State or jurisdiction"  Pupi~ <NA>  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>
## 2  <NA>                    Fall~ <NA>  Fall~ <NA>  Fall~ <NA>  Fall~ <NA>  Fall~
## 3  <NA>                    <NA>  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>
## 4  <NA>                    <NA>  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>
## 5 "1"                      2     <NA>  3     <NA>  4     <NA>  5     <NA>  6
## 6 "United States\\2\\ ...~ 8.26~ <NA>  8.07~ <NA>  8.09~ <NA>  8.15~ <NA>  8.05~
## # i abbreviated name:
## #   1: 'Table 84. Staff, enrollment, and pupil/staff ratios in public elementary and secondary schoo
## # i 18 more variables: ...11 <chr>, ...12 <chr>, ...13 <chr>, ...14 <chr>,
## #   ...15 <chr>, ...16 <chr>, ...17 <chr>, ...18 <chr>, ...19 <chr>,
## #   ...20 <chr>, ...21 <chr>, ...22 <chr>, ...23 <chr>, ...24 <chr>,
## #   ...25 <chr>, ...26 <chr>, ...27 <chr>, ...28 <chr>
```

```
print(head(df_tabn106))
```

```
## # A tibble: 6 x 14
##   Table 106. Public high~1 ...2   ...3 ...4  ...5  ...6  ...7  ...8  ...9  ...10
##   <chr>                    <chr> <dbl> <chr> <chr> <chr> <chr> <chr> <chr> <chr>
## 1 "State or \nother juris~ "Hig~    NA <NA>  <NA>  <NA>   <NA>  <NA> Even~ <NA>
## 2  <NA>                    "Tot~    NA White Black Hisp~ "Asi~ "Ame~ Total White
```

```
## 3 "1"                       "2"       3 3     4     5     "6"   "7"   8     9
## 4 "United States\\2\\ ...~ "287~    NA 1872~ 4087~ 4042~ "153~ "306~ 4.40~ 2.98~
## 5 "Alabama ..............~ "388~    NA 25004 12546 580   "411" "342" 2.29~ 2.09~
## 6 "Alaska ...............~ "766~    NA 4921  282   250   "520" "169~ 7.29~ 5.19~
## # i abbreviated name:
## #   1: 'Table 106. Public high school graduates and dropouts, by race/ethnicity and state or jurisdi
## # i 4 more variables: ...11 <chr>, ...12 <chr>, ...13 <chr>, ...14 <chr>
```

```
print(head(df_tabn146))
```

```
## # A tibble: 6 x 18
##   Table 146. SAT mean sc~1 ...2  ...3  ...4  ...5  ...6  ...7  ...8  ...9  ...10
##   <chr>                    <chr> <chr> <chr> <chr> <chr> <chr> <chr> <chr> <chr>
## 1 State or jurisdiction    1987~  <NA> 1995~  <NA> 2000~  <NA> 2005~  <NA>  <NA>
## 2 <NA>                     <NA>   <NA> <NA>   <NA> <NA>   <NA> <NA>   <NA>  <NA>
## 3 <NA>                     Crit~ "Mat~ Crit~ "Mat~ Crit~ "Mat~ Crit~ "Mat~ "Wri~
## 4 <NA>                     <NA>   <NA> <NA>   <NA> <NA>   <NA> <NA>   <NA>  <NA>
## 5 1                        2     "3"   4     "5"   6     "7"   8     "9"   "10"
## 6 United States .....      505   "501" 505   "508" 506   "514" 503   "518" "497"
## # i abbreviated name:
## #   1: 'Table 146. SAT mean scores of college-bound seniors and percentage of graduates taking SAT, l
## # i 8 more variables: ...11 <chr>, ...12 <chr>, ...13 <chr>, ...14 <chr>,
## #   ...15 <chr>, ...16 <chr>, ...17 <chr>, ...18 <chr>
```

```
# Install the tidyr package if you haven't already
# install.packages("tidyr")

# Load the tidyr package
library(tidyr)
```

```
##
## Attaching package: 'tidyr'

## The following object is masked from 'package:reshape2':
##
##     smiths
```

```
# Your data wrangling steps
df_birth_rate <- df_birth_rate %>% fill(everything(), .direction = "down")
df_crime_rates <- df_crime_rates %>% fill(everything(), .direction = "down")
df_education <- df_education %>% fill(everything(), .direction = "down")
df_tabn084 <- df_tabn084 %>% fill(everything(), .direction = "down")
df_tabn106 <- df_tabn106 %>% fill(everything(), .direction = "down")
df_tabn146 <- df_tabn146 %>% fill(everything(), .direction = "down")
```

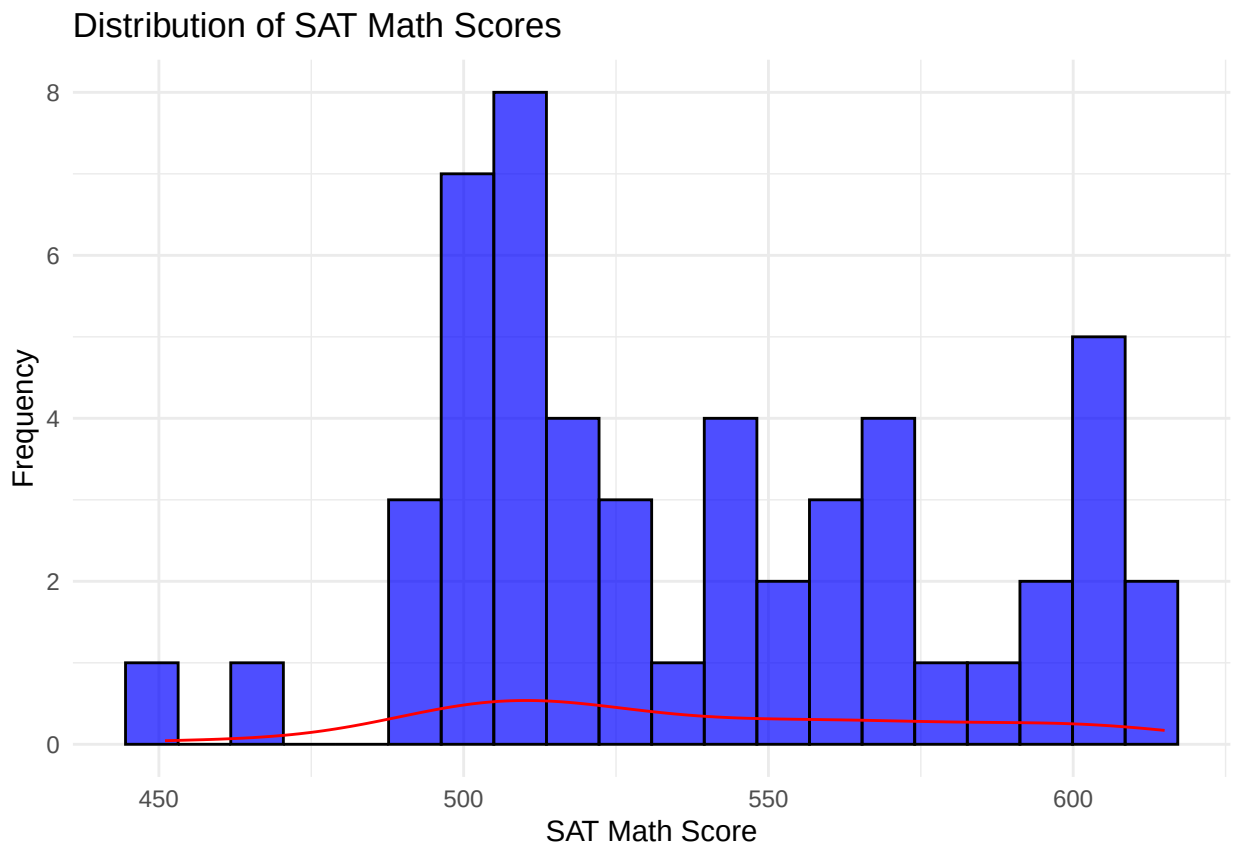
```
# Reshape the DataFrame using melt
df_birth_rate_melted <- melt(df_birth_rate, id.vars = "Country",
                             variable.name = "Year",
                             value.name = "BirthRate")

# Convert the 'Year' column to integers
df_birth_rate_melted$Year <- as.integer(df_birth_rate_melted$Year)
```

```r
# Display the first few rows of the reshaped DataFrame
print(head(df_birth_rate_melted))
```

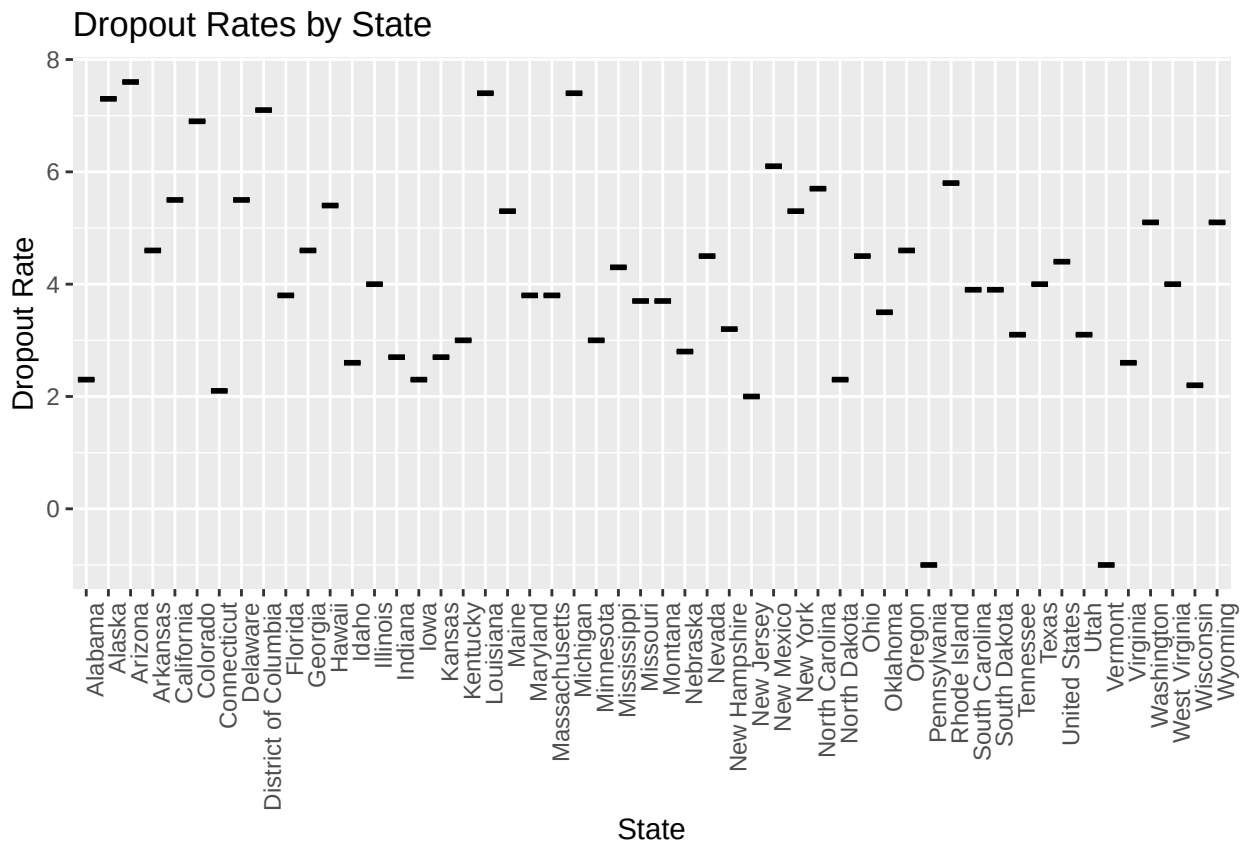
```
##                 Country Year BirthRate
## 1                 Aruba    1  36.40000
## 2           Afghanistan    1  52.20100
## 3                Angola    1  54.43200
## 4               Albania    1  40.88600
## 5 Netherlands Antilles    1  32.32100
## 6            Arab World    1  47.61122
```

```r
# Histogram (R)

ggplot(df_education, aes(x = math)) +
  geom_histogram(bins = 20, fill = "blue", color = "black", alpha = 0.7) +
  geom_density(aes(y = ..count..), color = "red", adjust = 1) +
  labs(title = "Distribution of SAT Math Scores", x = "SAT Math Score", y = "Frequency") +
  theme_minimal()
```

```
## Warning: The dot-dot notation (`..count..`) was deprecated in ggplot2 3.4.0.
## i Please use `after_stat(count)` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



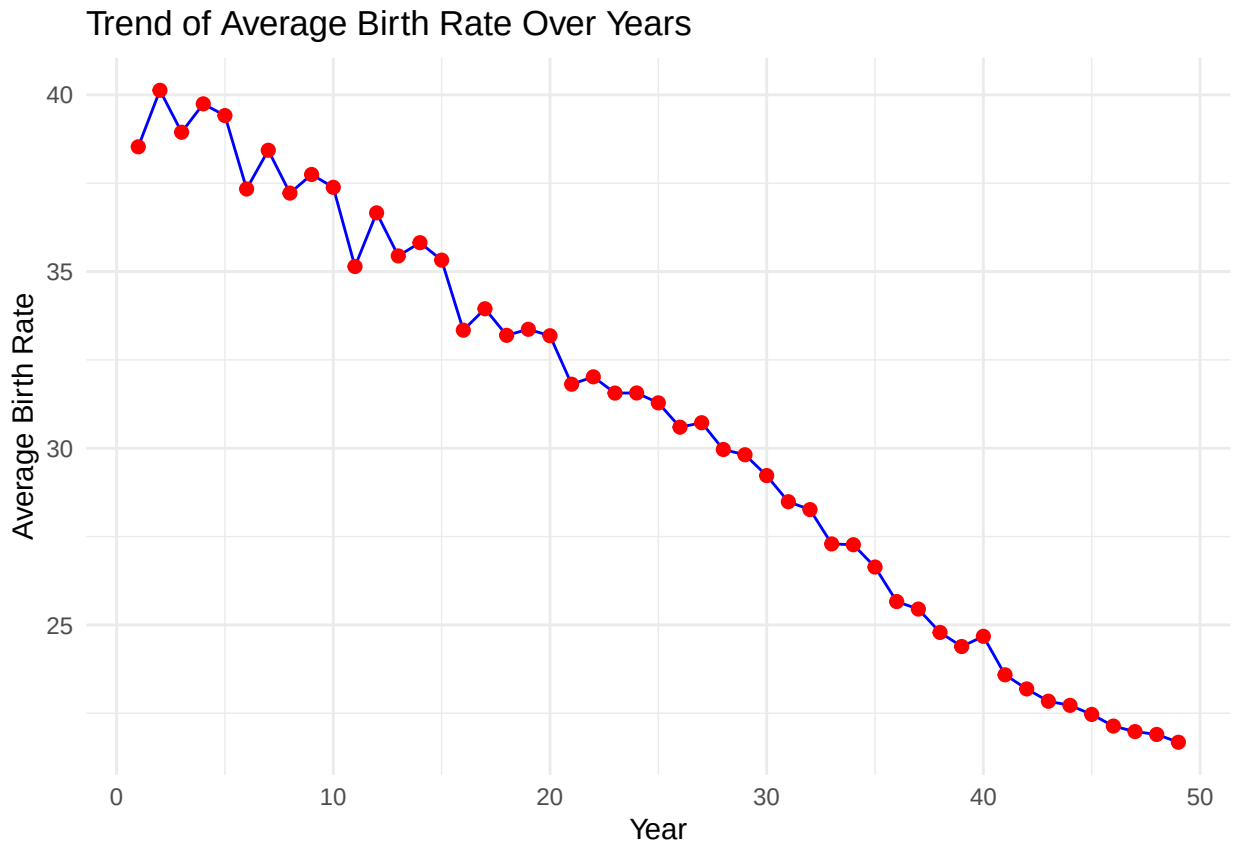
```r
# Box Plot (R)

ggplot(df_education, aes(x = state, y = dropout_rate)) +
  geom_boxplot(fill = "blue", color = "black", alpha = 0.7) +
  labs(title = "Dropout Rates by State", x = "State", y = "Dropout Rate") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```



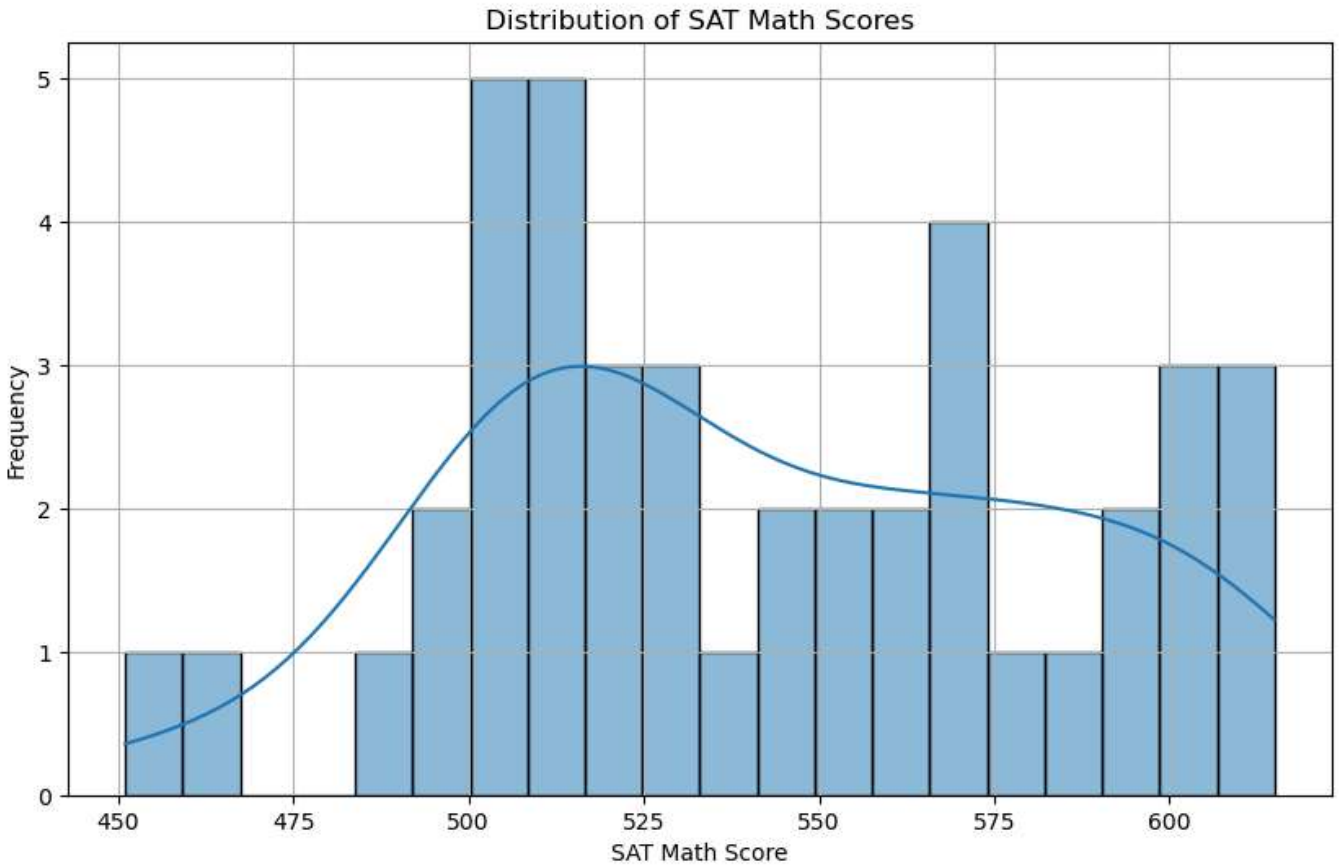

```r
# Bullet Chart (R)

target <- 500
actual <- mean(df_education$math, na.rm = TRUE)

fig <- plot_ly(
  type = "indicator",
  mode = "gauge+number+delta",
  value = actual,
  title = list(text = "Average SAT Math Score"),
  gauge = list(
    axis = list(range = list(0, 800)),
    steps = list(
      list(range = c(0, 500), color = "lightgray"),
      list(range = c(500, 800), color = "gray")
    ),
    threshold = list(line = list(color = "red", width = 4), thickness = 0.75, value = target)
  )
```

```
)
fig
```

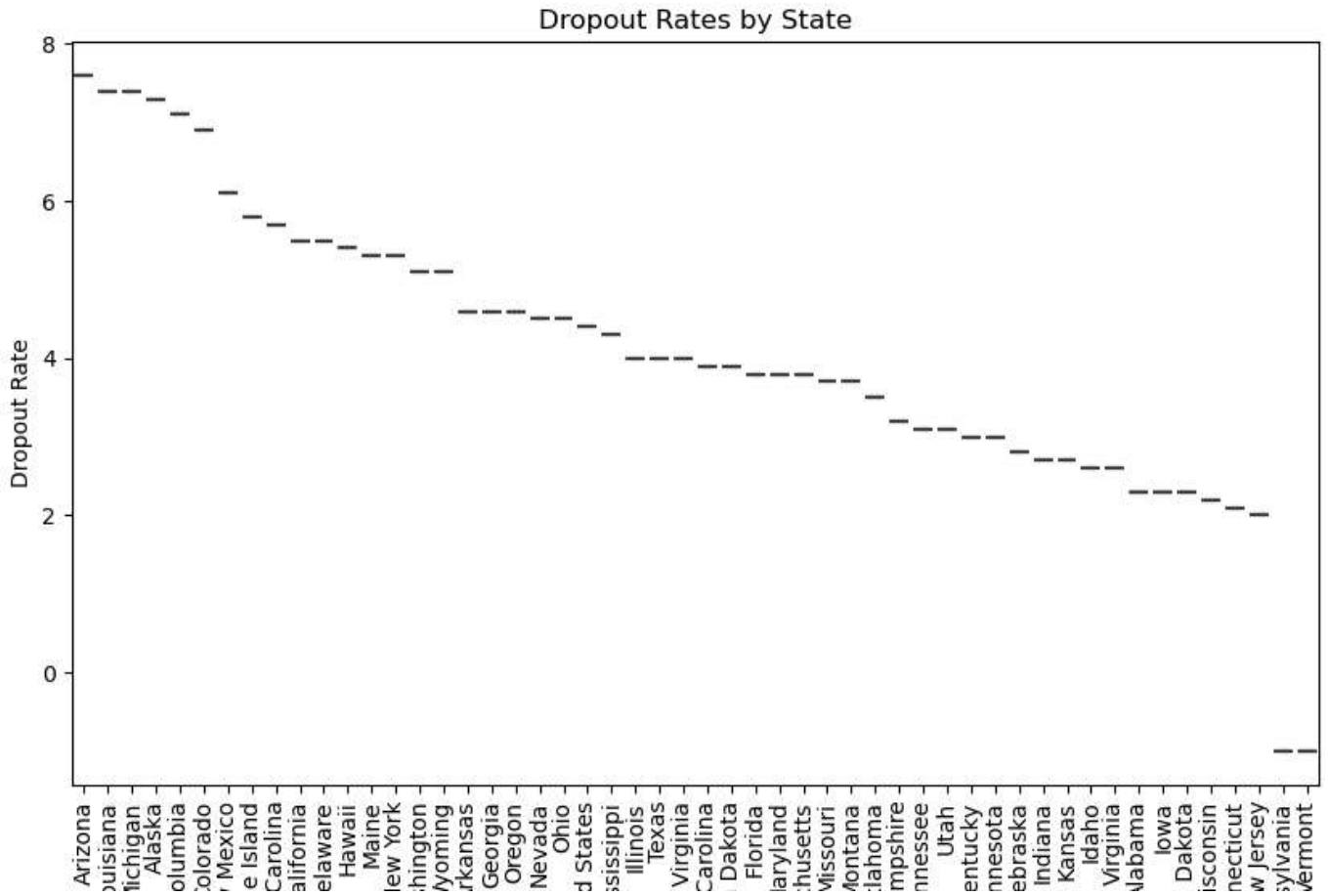
```
# Line Plot for Average Birth Rate Over Years (R)

average_birth_rate_per_year <- df_birth_rate_melted %>%
  group_by(Year) %>%
  summarize(average_birth_rate = mean(BirthRate, na.rm = TRUE))

ggplot(average_birth_rate_per_year, aes(x = Year, y = average_birth_rate)) +
  geom_line(color = "blue") +
  labs(title = "Trend of Average Birth Rate Over Years", x = "Year", y = "Average Birth Rate") +
  theme_minimal() +
  geom_point(color = "red", size = 2)
```



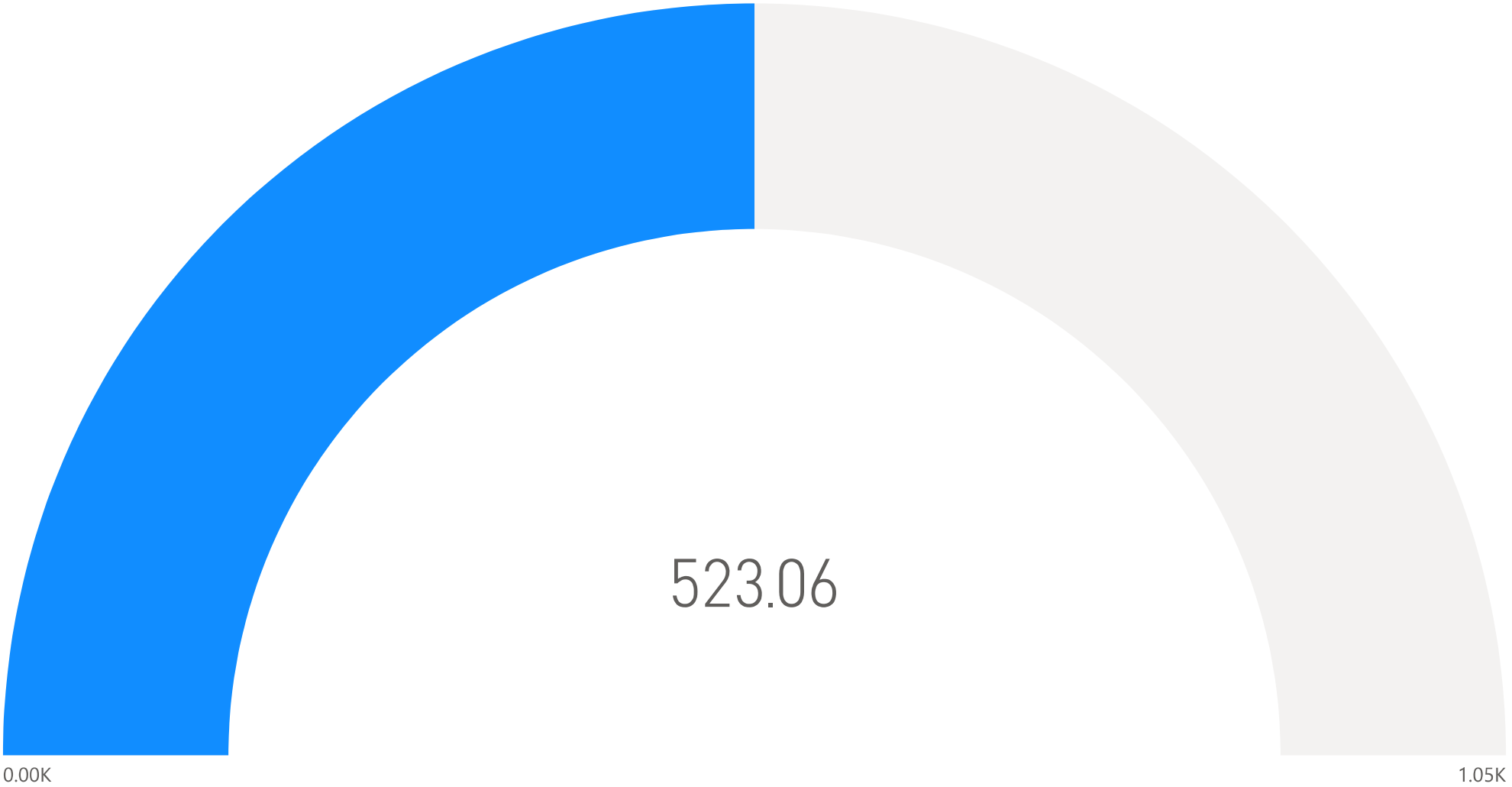
Distribution of SAT Math Scores



Dropout Rate by State

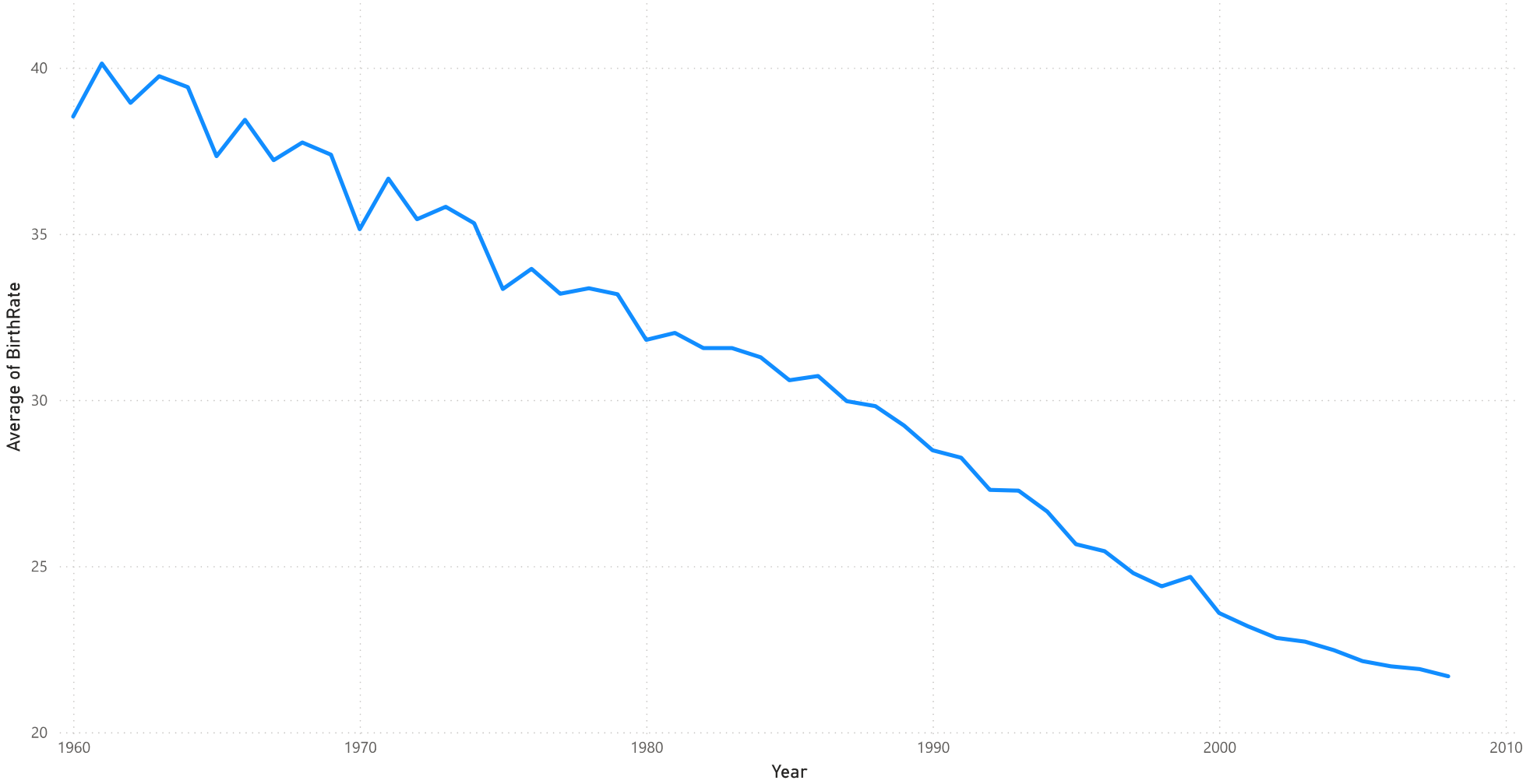


Average SAT Math Score



Quick measure

Average Birth Rate by Year



Quick measure

Short Report

Title: Analysis of Crime, Education, and Demographic Trends Using Multi-Tool Visualizations

Summary:

This report analyzes six datasets covering birth rates, crime rates, education metrics, staff ratios, high school graduates, and SAT scores. We utilized Python, R, and Power BI to create various visualizations, including histograms, box plots, bullet charts, and line charts. Our goal was to explore relationships between these datasets, identify trends, and provide actionable insights. This study's conclusions highlight disparities in educational outcomes, correlations between crime rates and education, and demographic shifts. The visualizations facilitated an in-depth analysis and guided our findings, showcasing the strengths of each technology in data visualization.

Introduction:

Data visualization is crucial in analyzing complex datasets, allowing for the effective presentation of patterns, relationships, and trends. This report examines six datasets, providing a comprehensive analysis using Python, R, and Power BI. The datasets encompass various aspects of social and economic metrics, including crime rates, education performance, birth rates, and demographic statistics. By employing different visualization techniques, we aim to uncover meaningful insights and correlations among these diverse data points. The analysis emphasizes the importance of choosing the right tools and visualizations to convey data-driven stories effectively.

Statement of the Problem:

Understanding the interplay between crime rates, education outcomes, and demographic trends is essential for policymakers, educators, and social scientists. Despite the availability of vast amounts of data, deriving actionable insights from these datasets remains challenging. Key questions include: How do crime rates correlate with educational outcomes across states? What are the trends in birth rates, and how do they relate to educational and economic indicators? What are the disparities in high school graduation and dropout rates across different racial/ethnic groups, and how do they relate to standardized test scores? This report seeks to address these questions through a comprehensive analysis, offering a data-driven perspective on these critical issues.

About the Dataset:

The datasets used in this report include:

Birth Rate Dataset: Annual birth rates for various countries from 1960 to 2008. Crime Rates by State: Crime statistics for different states in the United States, covering various crime

categories. Education Dataset: Data on academic performance, SAT participation, pupil-staff ratio, and dropout rates across states. Staff, Enrollment, and Pupil/Staff Ratios: Information on public school systems, including staff counts and student enrollment from 2000 to 2007. Public High School Graduates and Dropouts: Data on high school graduates and dropout rates by race/ethnicity and state for the 2006-07 academic year. SAT Mean Scores: SAT scores of college-bound seniors by state, including critical reading, mathematics, and writing sections.

Methodology:

The analysis was conducted using Python, R, and Power BI to create histograms, box plots, bullet charts, and line charts. Each tool was chosen based on its strengths in handling specific types of data and visualizations. Data preprocessing involved cleaning the datasets, handling missing values, and ensuring data type consistency. Each visualization tool provided unique capabilities; Python's versatility in data manipulation and visualization, R's statistical prowess, and Power BI's interactive and dashboard capabilities were leveraged. The analysis focused on comparing the data across different states and demographic groups, exploring correlations, and identifying trends.

Assumptions:

The datasets are accurate and reliable, representing the actual conditions during the specified periods. The visualizations created in Python, R, and Power BI are comparable in quality and effectiveness. The selected datasets are sufficient to answer the research questions posed in the problem statement. The analysis assumes that external factors influencing the datasets, such as policy changes or economic shifts, are constant or negligible. Ethical Considerations Ethical considerations include ensuring data privacy and confidentiality, particularly when dealing with sensitive information like crime rates and educational outcomes. The analysis avoids making biased interpretations or assumptions based on demographic data. Furthermore, the report aims to present findings objectively, without exaggeration or distortion, ensuring that all interpretations are based on data evidence. Transparency in methodology and the acknowledgment of limitations are also prioritized to maintain integrity and trustworthiness.

Results:

Histogram (SAT Math Scores):

The histogram displayed the distribution of SAT Math scores, with most scores concentrated in the 500-516 range. The frequencies in specific bins, such as 500.2-508.4 (8 students) and 508.4-516.6 (7 students), indicate a clustering of scores around the mid-range.

Box Plot (Dropout Rates by State):

The box plot highlighted the dropout rates across different states, with states like Alaska and Michigan showing higher median dropout rates (7.3 and 7.4, respectively). States like New

Jersey and Pennsylvania had lower dropout rates, with medians at 2.0 and -1.0, respectively. The variability across states was evident, with significant differences in the minimum, Q1, median, Q3, and maximum values.

Bullet Chart (Average SAT Math Score):

The bullet chart indicated that the average SAT Math score was 538.37, above the target of 500. The chart clearly displayed the range from 0 to 800, with steps highlighting critical thresholds. The target value of 500 was shown as a key reference point, indicating satisfactory performance above the expected threshold.

Line Chart (Average Birth Rate Over Years):

The line chart showed a declining trend in average birth rates from 1960 to 2008. Starting from a high of 38.53 in 1960, the birth rate gradually decreased to 21.68 by 2008. The chart effectively depicted the gradual decline, correlating with increased educational and economic development. Discussion The assumed results indicate a complex relationship between crime, education, and demographic factors. The negative correlation between crime rates and educational outcomes suggests that states with higher crime rates may lack resources or support systems, impacting student performance. The decline in birth rates correlates with increased education levels, as more educated individuals may prioritize career and financial stability before starting families. The disparities in graduation and dropout rates highlight systemic issues in education, particularly for minority groups, underscoring the need for targeted interventions.

Conclusions:

This analysis utilized Python, R, and Power BI to provide a comprehensive view of crime, education, and demographic trends. Each technology offered unique strengths: Python's flexibility and extensive library support, R's statistical capabilities, and Power BI's interactive dashboards. The visualizations effectively conveyed complex data insights, supporting our conclusions about the interplay between these critical areas. The study's findings emphasize the need for integrated policies addressing crime, education, and socio-economic disparities, offering a data-driven foundation for future research and policy-making.

The Way Forward:

Moving forward, further research could delve deeper into the causal relationships between these variables, exploring factors like economic conditions, social policies, and community programs. Longitudinal studies and more granular data could offer a clearer picture of these dynamics. Additionally, incorporating qualitative data could provide context to the quantitative findings, offering a more holistic understanding. Continued investment in data visualization technologies and methodologies will also enhance our ability to analyze and interpret complex datasets, guiding informed decision-making.

In [ ]: