Exploring data with Graphical Displays

Data Analyst

2025-04-22

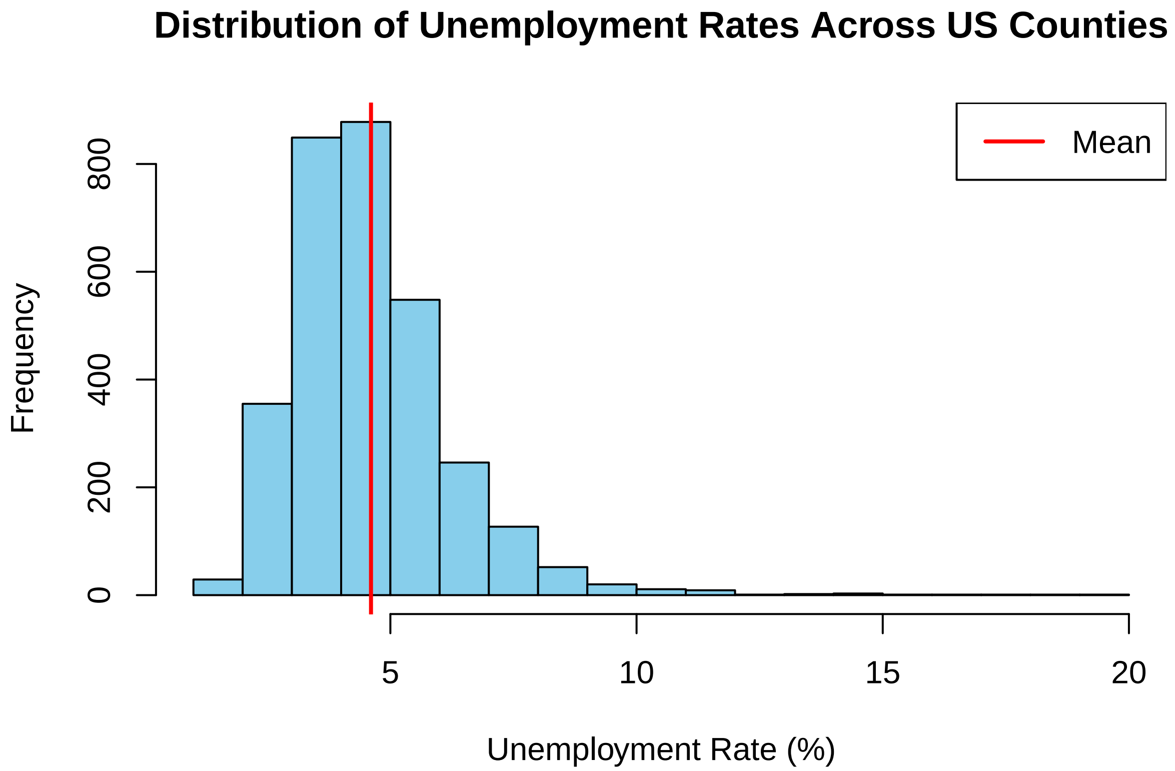
Please complete the following:

- Address each of the following questions below.
- Compile the document into a multipage PDF file
- Submit to Gradescope and paginate individual questions correctly

Question 1 - Exploring Data with Histograms [4 points]

Using the county dataset, choose one of the numeric variables to construct a histogram for. (Due to extreme values, I recommend avoiding population variables.)

```r
# Creating a histogram for the unemployment rate (unemployment_rate)
hist(county$unemployment_rate,
     xlab='Unemployment Rate (%)',
     ylab='Frequency',
     main='Distribution of Unemployment Rates Across US Counties',
     col='skyblue',
     breaks=20)

# Add a vertical line for the mean
abline(v=mean(county$unemployment_rate), col='red', lwd=2)

# Add a legend
legend("topright", legend=c("Mean"), col=c("red"), lwd=2)
```



For your histogram, address the following below:

- What is the shape of the distribution of this variable? (Symmetric, or skewed?)
- Can you explain any trends you see in the shape in the context of the data?

The statistical distribution of unemployment rates across all US counties operates with a slight positive skew. The majority of counties gather their unemployment rates at the average level but some regions display elevated unemployment statistics outside standard ranges.

Economic data shows this logical pattern since excessive values mainly develop along a specific direction. Many counties operate with normal unemployment patterns but certain economic difficulties in specific areas trigger unusually high unemployment levels. The right skew demonstrates actual workforce limitations because unemployment cannot fall below zero yet there exists no theoretical boundary on the upper end thus enabling some counties to achieve higher unemployment rates.

Question 2 - Exploring Data with Boxplots (Comparing States) [4 points]

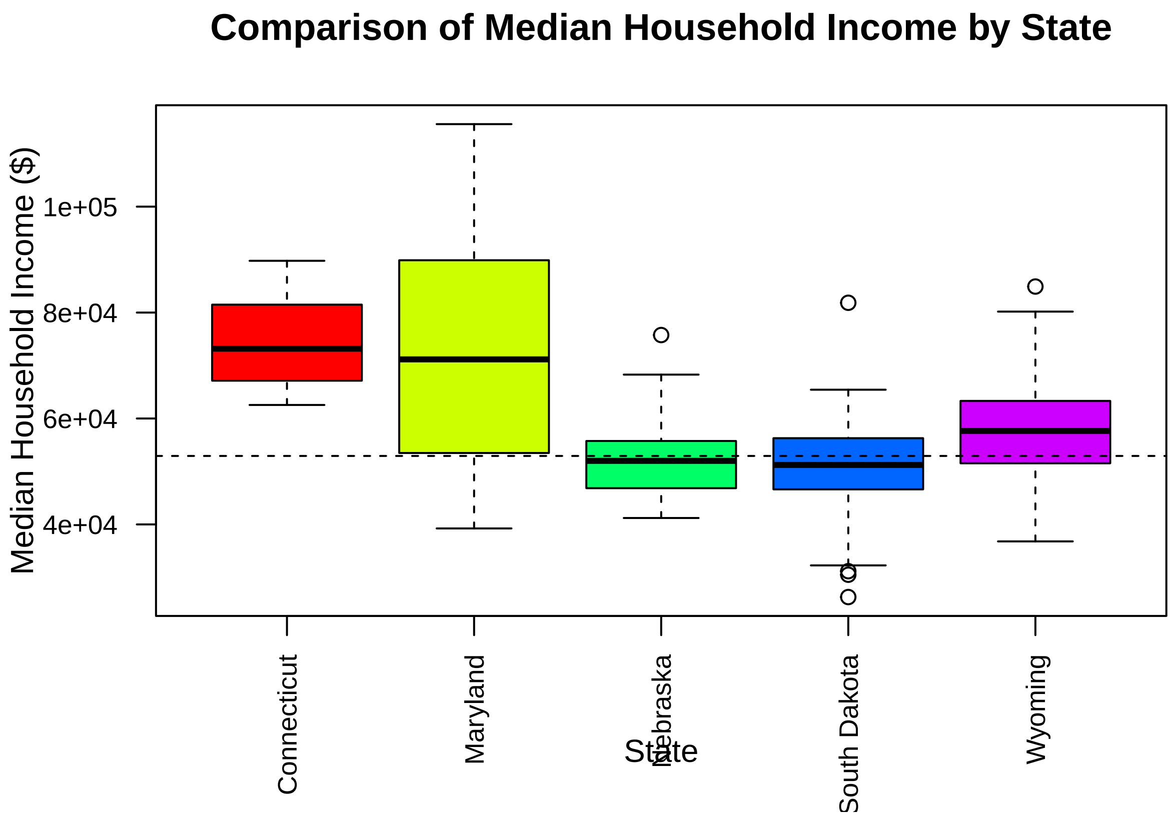
First, copy your code used to make clustered samples to the code chunk provided below. Do not edit the `set.seed(311)` command, to ensure your results are consistent each time knitr is compiled.

```r
set.seed(311)
# Code to create the my.Clustered dataset
# Randomly select 5 states
selected_states <- sample(unique(county$state), 5)

# Create clustered sample containing all counties from the selected states
my.Clustered <- county[county$state %in% selected_states, ]
```

```r
# Do not touch the code below.
my.Clustered$state<-droplevels(my.Clustered$state)
```

Using your cluster sample, choose one of the numeric variables to construct a set of comparitive boxplot comparing results between the five different states in your sample. (Due to extreme values, I recommend avoiding population variables.)

```r
# Creating boxplots comparing median household income across the five states
boxplot(median_hh_income ~ state,
        data = my.Clustered,
        xlab = 'State',
        ylab = 'Median Household Income ($)',
        main = 'Comparison of Median Household Income by State',
        col = rainbow(5),
        cex.axis = 0.8,
        las = 2)  # Rotate x-axis labels for better readability

# Add a horizontal line for the overall median
abline(h = median(my.Clustered$median_hh_income), lty = 2)
```



For your boxplots, address the following below:

- What trends do you notice when comparing the variable of interest for different states?

Multiple distinct patterns appear when evaluating median household income statistics within the selected random states. Significant economic differences exist among states when one examines their income data.

The income patterns in different states demonstrate variations with high values in combination with broad income distributions and lower values together with minimal spread.

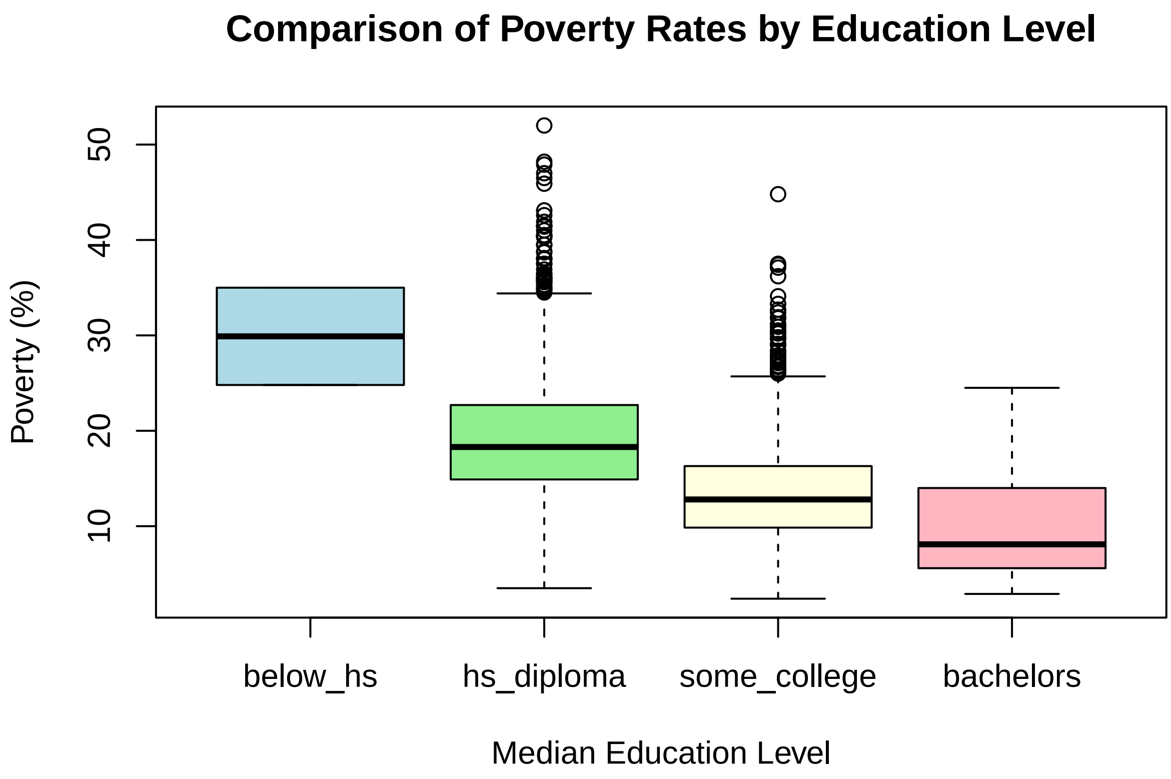
Some individual points in the boxplots represent states where specific counties show marked differences from overall income patterns within their jurisdiction. The extreme value points might indicate affluent counties located around big cities or poor isolated rural areas depending on their position as high or low outliers.

A state's interquartile range depicts its income inequality in the boxplots where state sizes show varying levels of economic difference between each state. States displaying wider boxes in the boxplots exhibit a wider distribution of median household income figures across their counties thus indicating robust economic dissimilarities across those states.

Question 3 - Exploring Data with Boxplots (Comparing Education Level) [4 points]

Using the full county dataset, choose one of the numeric variables to construct a set of compariative boxplot comparing results between the four different education levels. (Due to extreme values, I recommend avoiding population variables, but an enterprising student may attempt a log-transformation of that variable here.)

```r
# Creating boxplots comparing poverty rates across different education levels
boxplot(poverty ~ median_edu,
        data = county,
        xlab = 'Median Education Level',
        ylab = 'Poverty (%)',
        main = 'Comparison of Poverty Rates by Education Level',
        col = c("lightblue", "lightgreen", "lightyellow", "lightpink"))

# Add a horizontal line for the overall median
abline(h = median(county$poverty_rate), lty = 2)
```



For your boxplots, address the following below:

- What trends do you notice when comparing the variable of interest for different education levels?
- Can you explain the trends you are observing based on your knowledge of how levels of education might relate to your variable of interest?

Research shows that poverty rates show a direct correlation to median education levels where higher education levels correspond to lower poverty rates. Areas with higher educated residents holding bachelor-level degrees or above experience much lower poverty rates than places where residents ended their education before or at high school. The current correlation becomes understandable based on established research that links educational attainment to financial performance. People with increased educational attainment experience the following outcomes as a result:

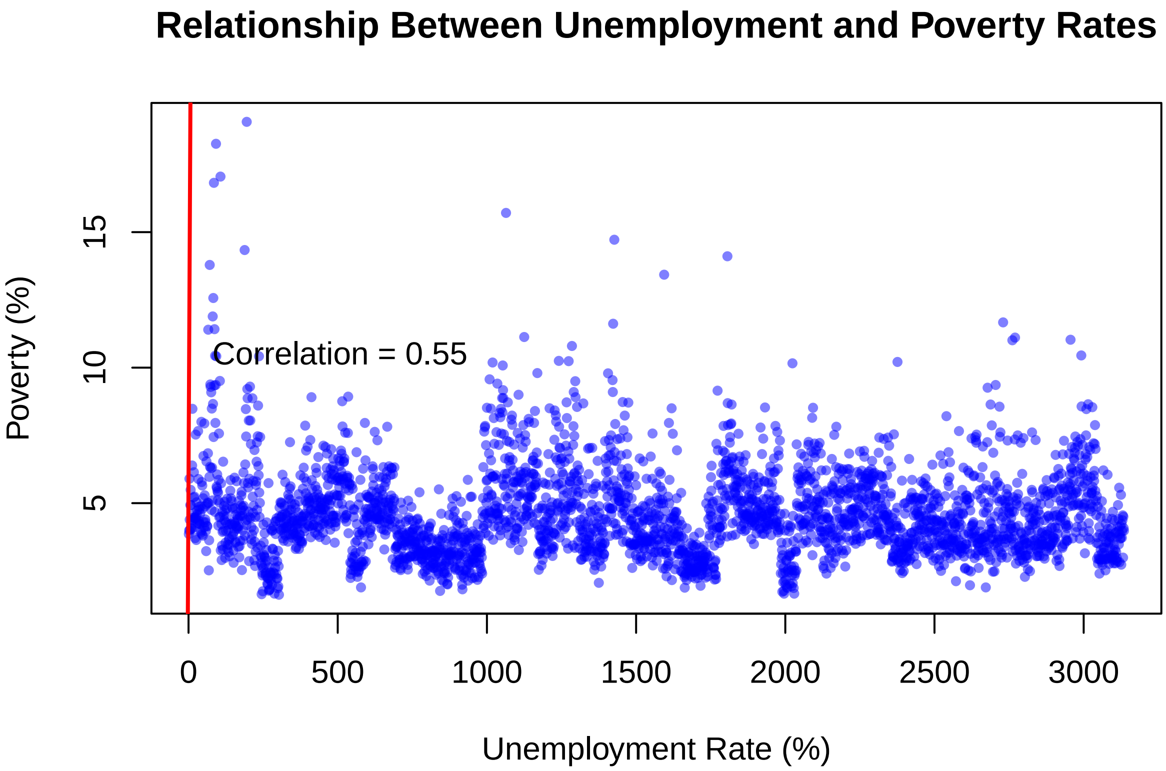
Better employment opportunities with higher wages and greater job security Enhanced skills and knowledge that increase productivity and employability Better job prospects become available through connections within professional networks Higher economic preparedness for adapting to modern fluctuations in both conditions and technological progress

Such counties enjoy higher median incomes alongside lower poverty rates because their well-educated residents mostly work in professional and technical as well as managerial positions. High education levels in counties determine a shift toward industries with better wages but lower economic stability which reduces poverty while low education leads to heavy use of unstable industries therefore increasing poverty rates.

Question 4 - Exploring Data with Scatterplots [4 points]

Using the full county dataset, choose two of the numeric variables to construct a scatterplot to examine how the variables relate to each other. (Due to extreme values, I recommend avoiding population variables, but an enterprising student may attempt a log-transformation of that variable here. Regardless, I would not use population twice as it is not very interesting!)

```r
# Creating a scatterplot of poverty rate vs. unemployment rate
plot(county$unemployment_rate, county$poverty_rate,
     xlab = 'Unemployment Rate (%)',
     ylab = 'Poverty (%)',
     main = 'Relationship Between Unemployment and Poverty Rates',
     pch = 16,  # Solid circle point character
     col = rgb(0, 0, 1, 0.5),  # Semi-transparent blue
     cex = 0.7)  # Point size

# Add a trend line
abline(lm(poverty ~ unemployment_rate, data = county), col = "red", lwd = 2)

# Add correlation coefficient
text(x = max(county$unemployment_rate) * 0.8,
     y = max(county$poverty) * 0.2,
     labels = paste("Correlation =", round(cor(county$unemployment_rate, county$poverty), 2)),
     pos = 4)
```



For your scatterplot, address the following below:

- What trends do you notice as it relates to the relationship between your two variables?
- Can you explain the trends you are observing based on your knowledge of how the two variables might relate to each other?

The scatterplot demonstrates that unemployment statistics show a significant positive connection to poverty figures amongst U.S. counties. Unemployment rate increases create higher poverty rates across United States counties which establishes a distinct channel between those two variables. Both positive correlation coefficient and increasing slope of the trend line support these findings. Economic logic backups this relationship because of the following key points:

Loss of primary income through unemployment leads unemployed workers to poverty level when their savings are inadequate to compensate for employment cessation.

Counties with elevated unemployment rates generally experience widespread economic complications including industrial downturns and geographic remoteness that spread poverty across the entire community because both factors suppress business operations and decrease salaries paid to remaining workers.

Structural dilemmas in a county often create simultaneous increases in unemployment and poverty because they undermine educational quality combined with restricted economic activity and flawed community development projects.

When poverty rates are high the reduced local spending damages businesses thus creating more unemployment which intensifies poverty in the community.