Exploring data with Graphical Displays

Anthony Wen

2024-01-24

Please complete the following:

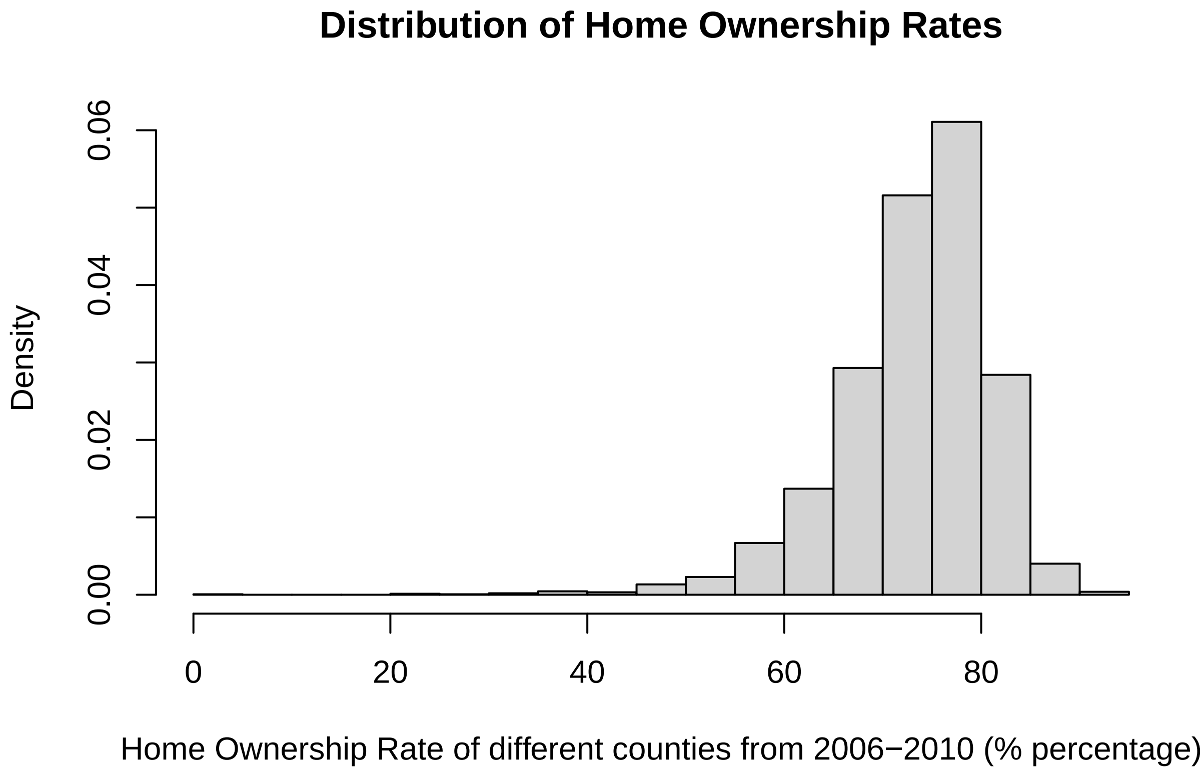
- Address each of the following questions below.
- Compile the document into a multipage PDF file
- Submit to Gradescope and paginate individual questions correctly

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
# Include code here to construct a histogram from the county dataset
#  Be sure to change the labels of the axes to be a proper graphical display!
# hist(NA, xlab='xlabel', ylab='ylabel', main='title')

hist(county$homeownership,freq = FALSE,breaks=20,
     xlab = "Home Ownership Rate of different counties from 2006-2010 (% percentage)",
     main = "Distribution of Home Ownership Rates")
```



For your histogram, address the following below:

- What is the shape of the distribution of this variable?
- Can you explain any trends you see in the shape in the context of the data?

The shape of this variable is most similar to a negative skew. For the 0-50% of home ownership range, the density is really different from the majority of the dataset from 60% and above. This means that the percentage of home ownership in different counties are mostly between 60% and 80% with the peak distribution between 70% to 80% (the mean is around 73%). So, there is a larger amount of counties that have high home ownership rate as opposed to low home ownership 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)
# Include your code from HW1 to create the my.Clustered dataset. Keep the added last line at the end.

clusters = sample(unique(county$state),5)
#clusters

clusters2 <- county[0,]
```

```r
for (cluster in clusters){
  temp<-county[county$state==cluster,]
  clusters2 <- rbind(clusters2,temp)
}
my.Clustered <- clusters2

#This should only give 5 total states
unique(my.Clustered$state)
```

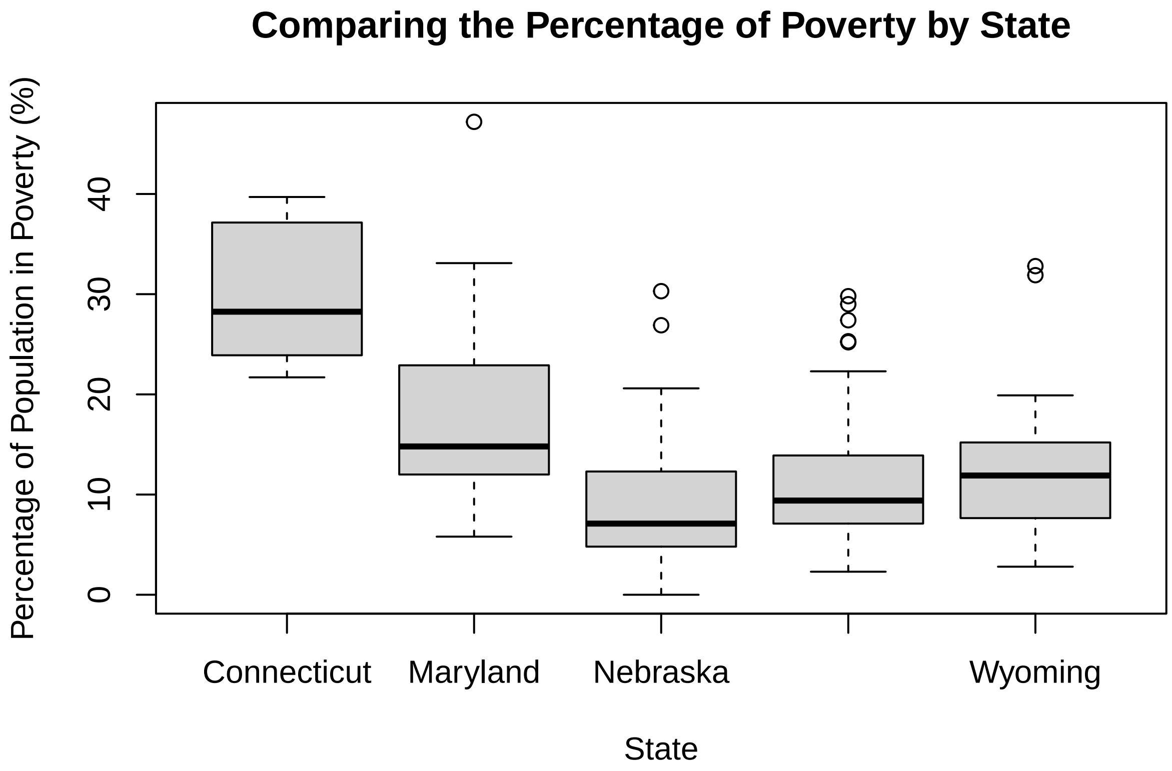
```
## [1] Wyoming      South Dakota Maryland     Nebraska     Connecticut
## 51 Levels: Alabama Alaska Arizona Arkansas California Colorado ... Wyoming
```

```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 compariative boxplot comparing results between the five different states in your sample. (Due to extreme values, I recommend avoiding population variables.)

```r
# Include code here to construct a compariative boxplot from the states in your
#  my.Clustered dataset
#  Be sure to change the labels of the axes to be a proper graphical display!
# boxplot(NA~my.Clustered$state, xlab='xlabel', ylab='ylabel', main='title')

boxplot(my.Clustered$multi_unit~my.Clustered$state,xlab = "State",
        ylab = "Percentage of Population in Poverty (%)",
        main= "Comparing the Percentage of Poverty by State")
```



For your boxplots, address the following below:

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

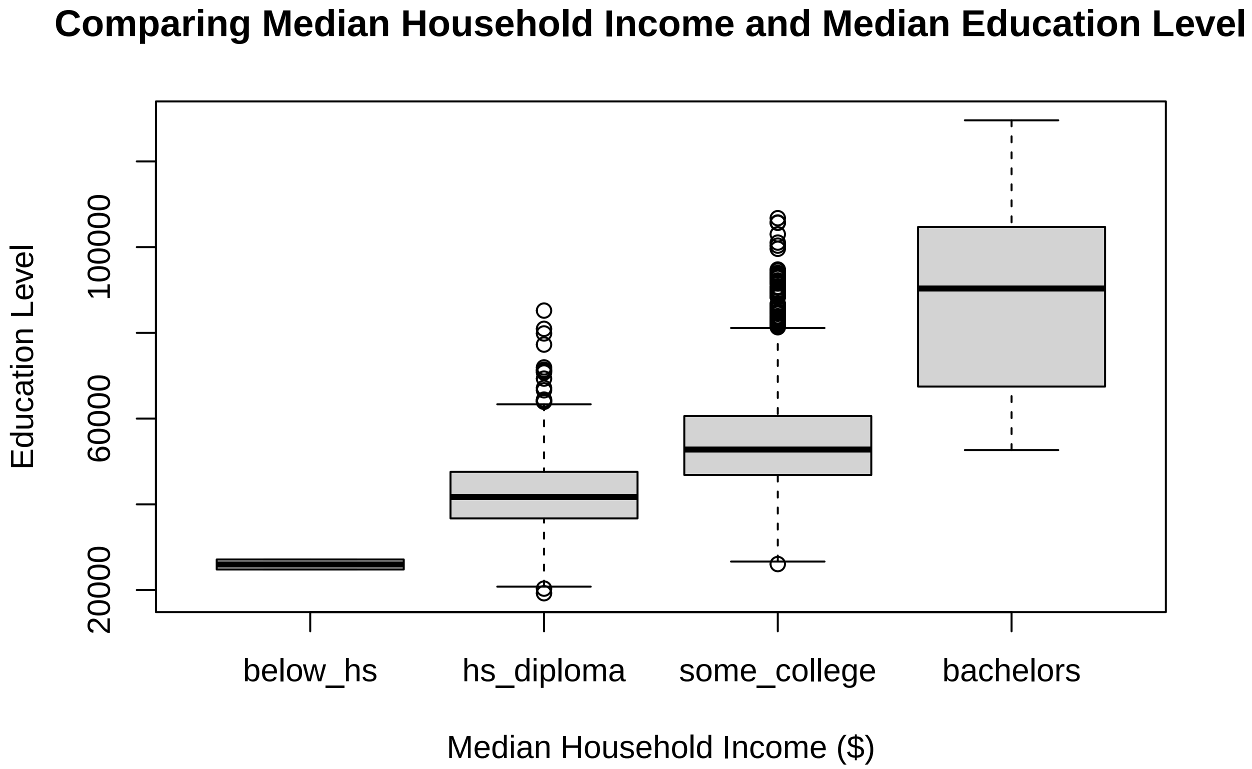
From the boxplot, we can see that Nebraska, South Dakota, Wyoming has counties that has a lower pencentage of population of counties with the majority of the counties below 15% poverty. In contrast, Connecticut has most of its counties having a higher population in poverty. It has over 20% of the population being in poverty. This also shows how the median values of Connecticut is the highest of all, with Wyoming, Nebraska, South Dakota having the lower of the 5. In addition, Connecticut also has the widest interquartile range (IQR). There are also a few outliers in the data for Nebraska and South Dakota which shows the counties with a percentage of poverty that is shockingly high compared to the other counties in the states. The boxplot for Maryland is also relatively symmetric compared to the others, which shows how the distribution for poverty percentage in the state is relatively symmetric.

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.)

```
# Include code here to construct a compariative boxplot from the education
# levels in the county dataset
#  Be sure to change the labels of the axes to be a proper graphical display!
# boxplot(NA~county$median_edu, xlab='xlabel', ylab='ylabel', main='title')
```

```
boxplot(county$median_hh_income~county$median_edu,xlab = "Median Household Income ($)",
        ylab = "Education Level",
        main="Comparing Median Household Income and Median Education Levels")
```



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?

From the plot, we can see a postive correlation between the median education level and median household income in the county. The higher the income, the higher the education level which suggest a right-skewed graph.The variability also seems to increase along this trend, the higher the education level (and income) the wider the IQR and the more outliers there might be (other than in the Bachelors catagory). This might be because the more money you have, the more resources you might be able to obtain which of course also includes education. Also, when people have a higher education level, they gain more experience and knowledge which suggests more job opportunities avalible to them. There will also have high salary levels (income).

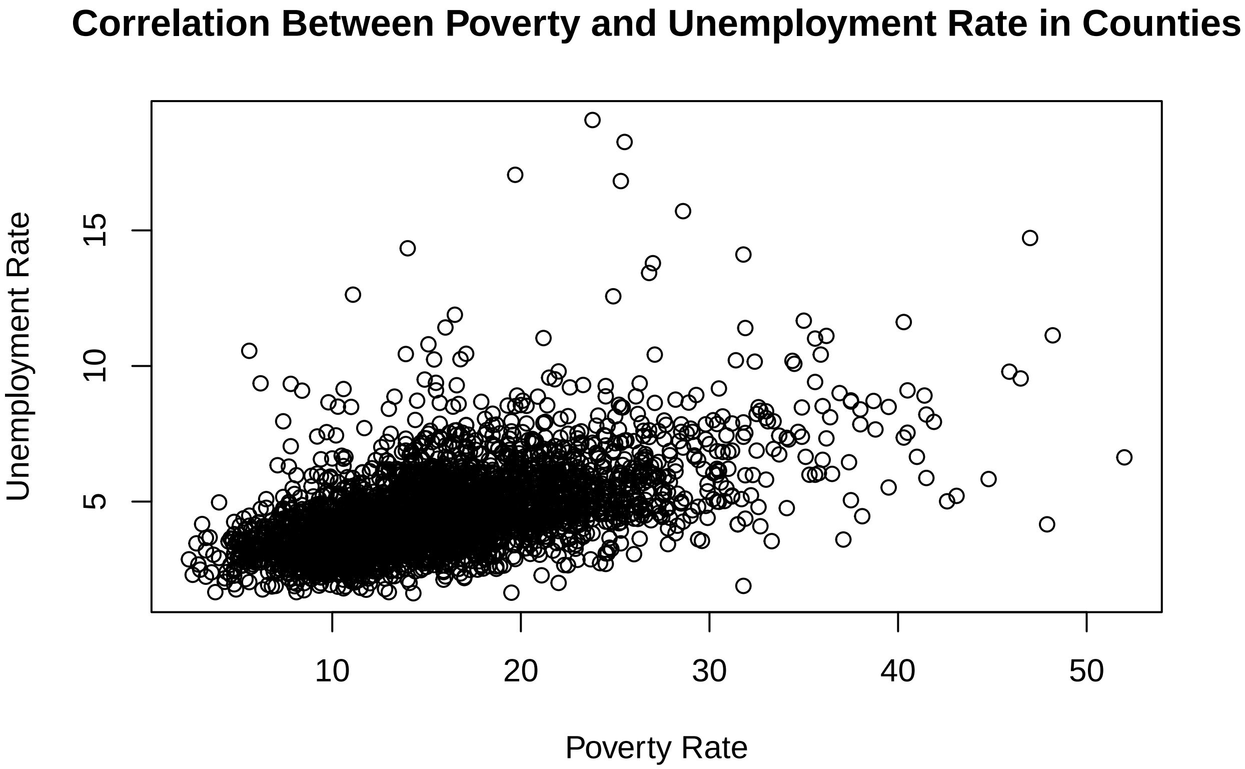
Question 4 - Exploring Data with Scatterplots [4 points]

Using the full county dataset, choose two of the numeric variables to construct a scaterplot 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!)

```
# Include code here to construct a scatterplot from the county dataset
#  Be sure to change the labels of the axes to be a proper graphical display!

# plot(county$Y~county$X, xlab='xlabel', ylab='ylabel', main='title')
# or
# plot(county$X, county$Y, xlab='xlabel', ylab='ylabel', main='title')

plot(county$unemployment_rate~county$poverty,xlab= 'Poverty Rate',
     ylab = 'Unemployment Rate',
     main='Correlation Between Poverty and Unemployment Rate in Counties')
```



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?

There seems to be a strong positive correlation between poverty rate and unemployment rate. As ppoverty rate inccreases the unemployment rate also increases. Most of the data seems to be on the lower rates of both poverty and unemployment meaning that most counties have lower rates of poverty and unemployment. However, as the rates of poverty and unemployment increase, the spread of data for the rates seems more spread out and not as dense as it was for lower rates. There are a few counties with particularly high poverty rates that stand out from the rest of the data, this might be a unique location within the sample data. Also, it makes sense because unemployment directly relates with how much money a person has. With no job, its more likely that a person does make a lot of money, so they would not have enough money, making them to be in poverty. Both factors could also be influenced by other things like education levels, access to jobs, local economy.