EDA Intro

Exploratory Data Analysis

Learning Outcomes

-

The following dataset is from the Current Population Survey in 1985. The following table summarizes the data.

Variable	Description
educ	Number of years of education
south	Indicator variable for living in a southern region: S = lives in south, NS = does not live in south
sex	Gender: M = male, F = female
exper	Number of years of work experience (inferred from age and education)
union	Indicator variable for union membership: Union or Not
wage	Wage (dollars per hour)
age	Age (years)
race	Race: W = white, NW = not white
sector	Sector of the economy: clerical, const (construction), management, manufacturing, professional, sales, service, other
married	Marital status: Married or Single

Vocabulary Review

1. What are the observational units?

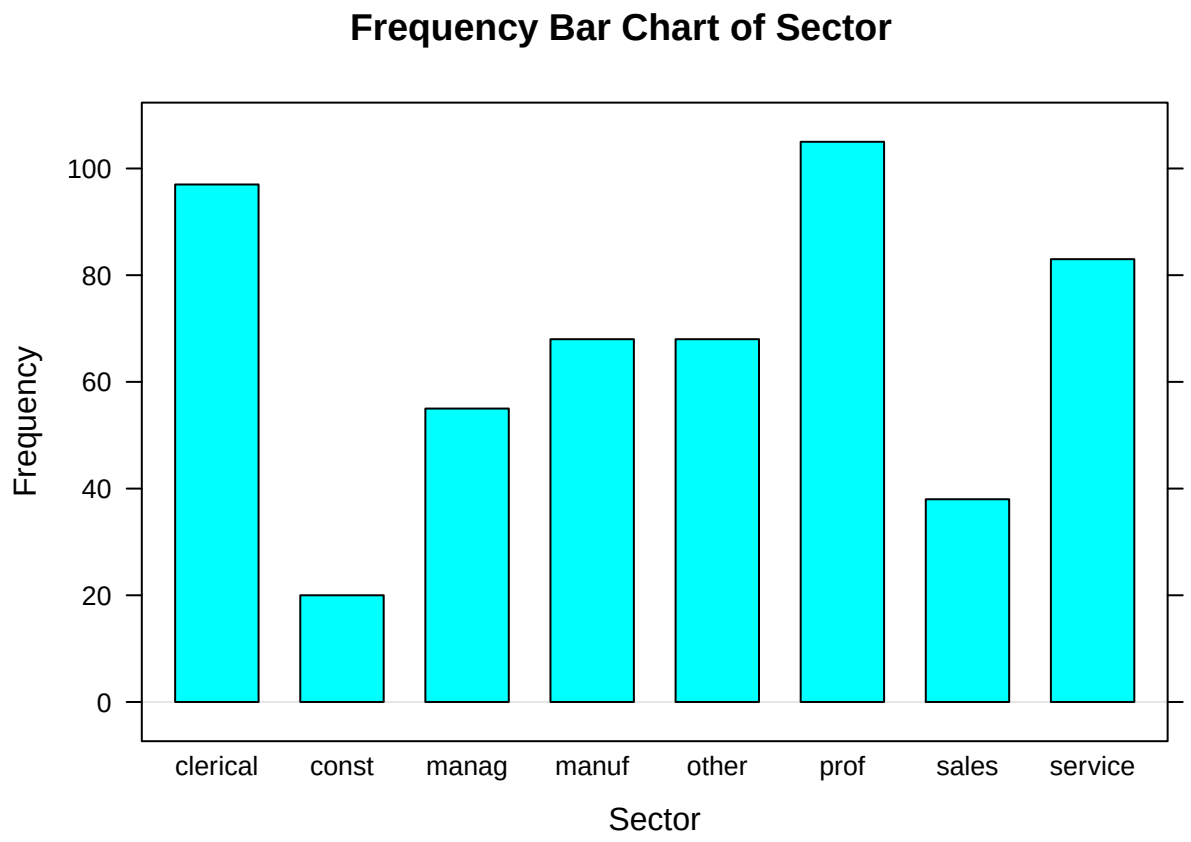
2. Which variables are categorical?

3. Which variables are quantitative?

A bar chart is used to plot a single categorical variable. We can plot the counts for each category in a frequency bar chart and the proportion in each categorcy in a relative frequency bar chart. Here we will create a bar chart of the variable sector.

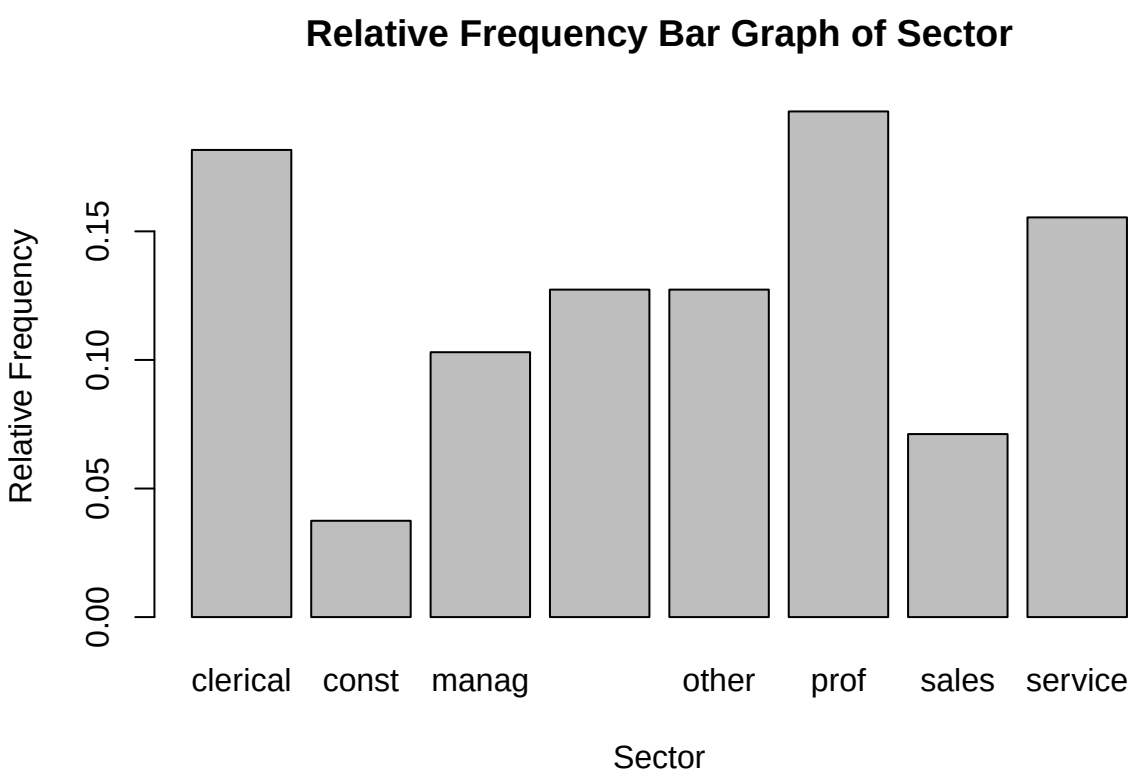
```
cps <- read.csv("../data/cps.csv") #This will read in the dataset
cps$sector <- factor(cps$sector) #When a variable is categorical you need to set up as a factor?????rew
cps$sex <- factor(cps$sex)
```

```
barchart(cps$sector, #This specifies the dataset and the variable
         horizontal = FALSE, #Turn the bars so they are vertical
         main = "Frequency Bar Chart of Sector", #Give your chart a title
         xlab = "Sector", #Label the x axis
         ylab = "Frequency", #Label the y axis
         ) #change the color of the bars
```



3. Which Sector has the largest number of people in it?

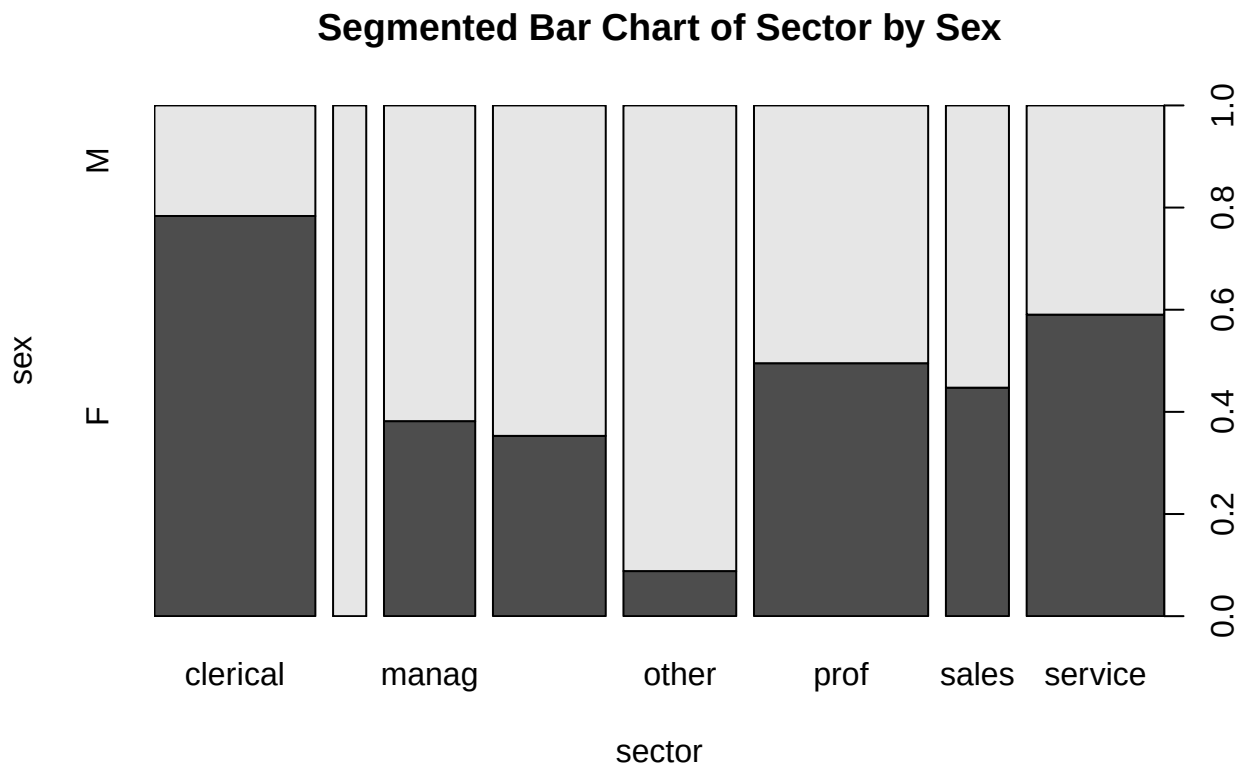
```
barplot(table(cps$sector)/nrow(cps), #divide the frequency counts by the total
        main = "Relative Frequency Bar Graph of Sector", #Give your chart a title
        xlab = "Sector", #Label the x axis
        ylab = "Relative Frequency", #Label the y axis
)
```



4. How does this plot differ from the plot above?

To visually display two categorical variables we will use a segmented bar chart.

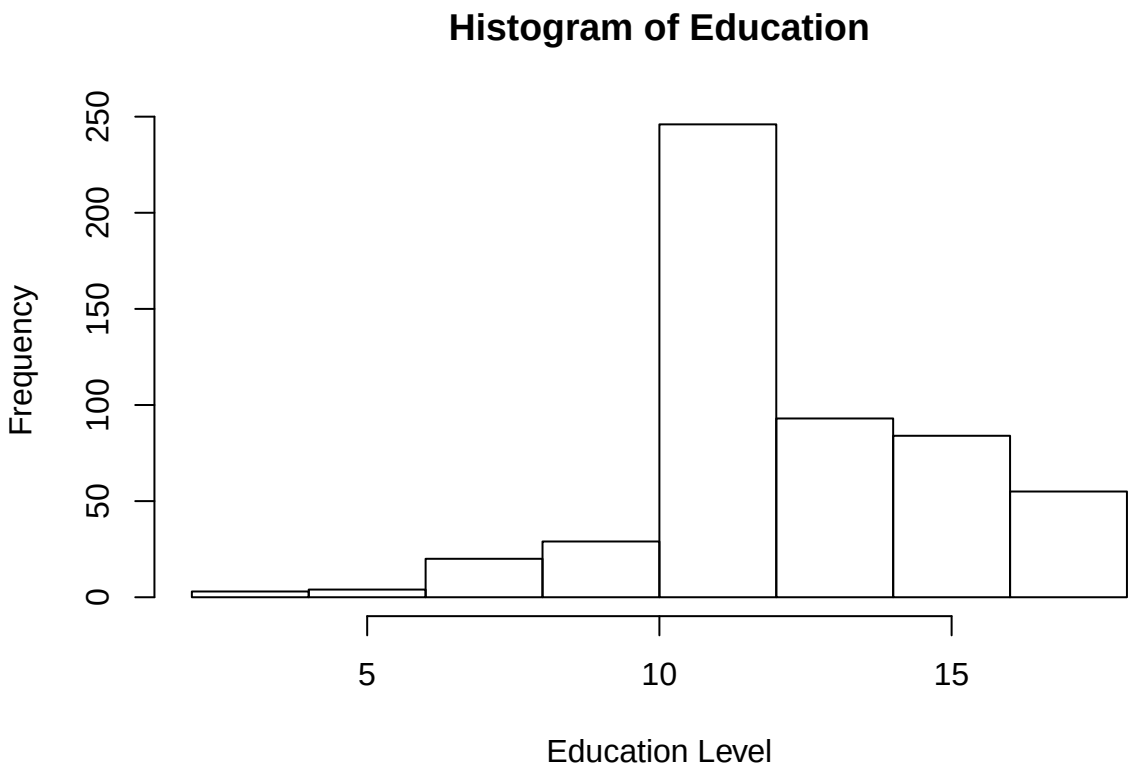
```
plot(sex~sector #response~explanatory allows us to plot two variables
     , data = cps, main="Segmented Bar Chart of Sector by Sex" #Make sure to title your graph
     )
```



5. Using the segmented bar chart, which sector has about the same proportion of males and females?

To plot quantitative variables, we can use a histogram or boxplot. To create a histogram the variable is broken into bins on a set width. Each bin plots the frequency of each .....$ We will create a histogram of the variable education.

```
hist(cps$educ, #dataset name and variable
     main = "Histogram of Education",
     xlab = "Education Level")
```



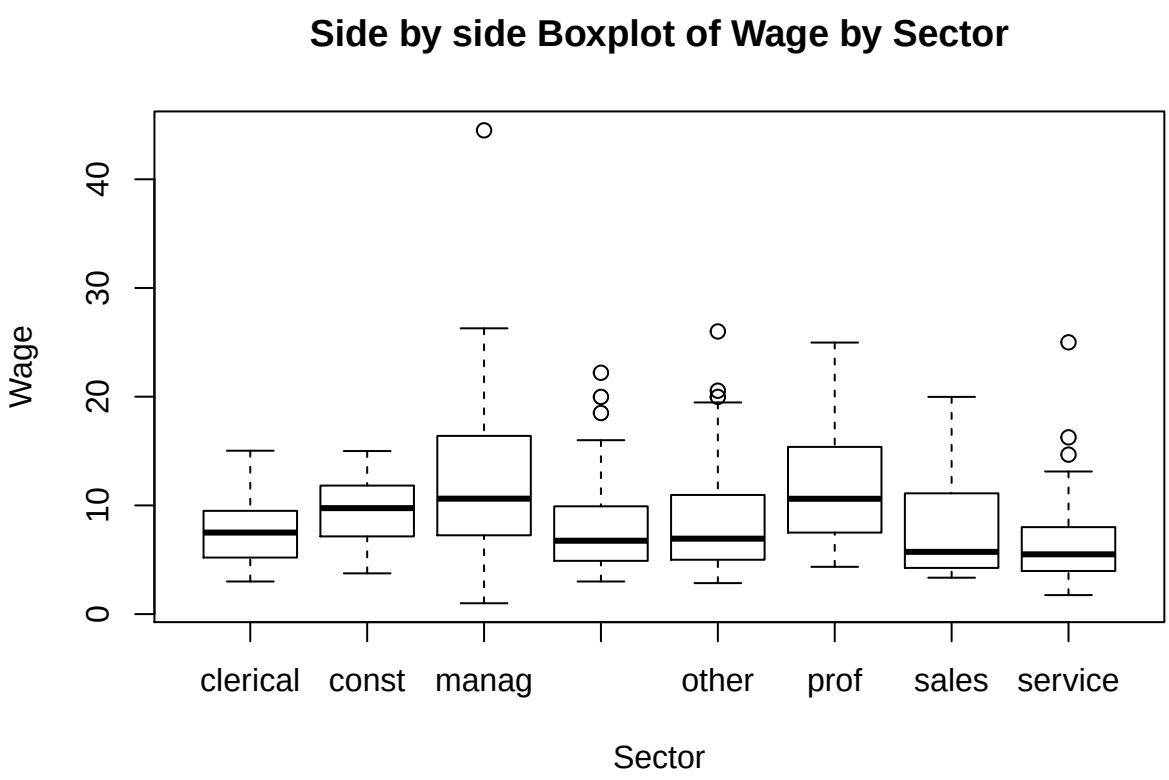
6. What is the width of each bin?

7. What is the most common level of Education?

Side by side plots are created from a single categorical and single quantitative variable. The boxplot is created using the five number summary: * Minimum value * Quartile 1 (Q1) - the value at the 25th percentile * Median - the value at the 50th percentile * Quaritle 3 (Q3) - the avlue at the 75th percentile * Maximum value Outliers are values less than $Q_1 - 1.5 * IQR$ and greater than $Q_3 + 1.5 * IQR$

The boxplot of wage by sector is plotted using the code below.

```
boxplot(wage~sector #response~explanatory
        , data=cps, main = "Side by side Boxplot of Wage by Sector",
        xlab = "Sector", ylab = "Wage")
```



8. Compare Service and Sales using the four characteristic to comparing distributions.

. . . Fill in the following table. . .