Exploratory Data Analysis

Exploratory Data Analysis

Learning Outcomes

- Identify and create appropriate summary statistics and plots given a data set or research question
- Interpret the following summary statistics in context: median, lower quartile, upper quartile, standard deviation, inter-quartile range, coefficient of determination, regression line slope
- Given a plot or set of plots, describe and compare the distribution(s) of a single quantitative variable (center, variability, shape, outliers)

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?

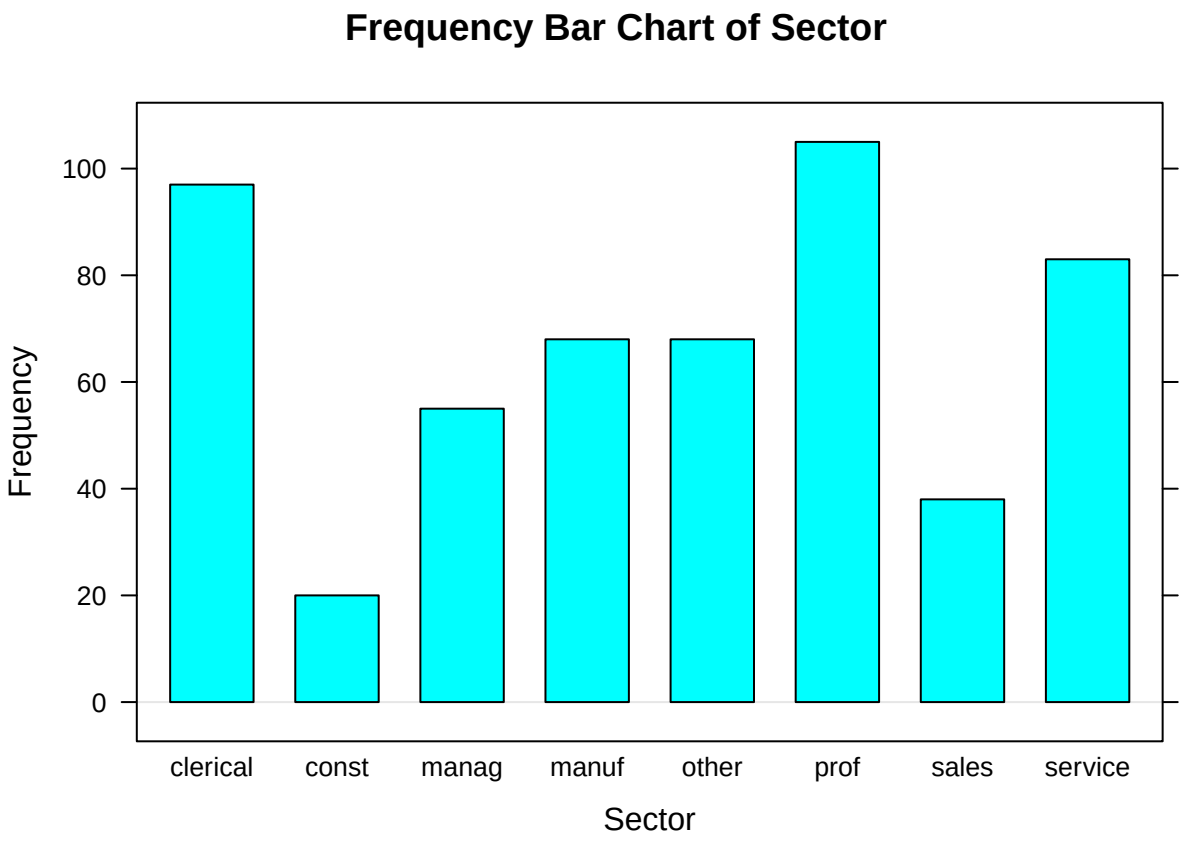
An important part of understanding data is to create visual pictures of what the data represents. In this activity we will create graphical representations of different types of data and different combinations of data.

Displaying a single categorical variable.

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. If we wanted to know how many people in our dataset were in each sector, we would 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)
```

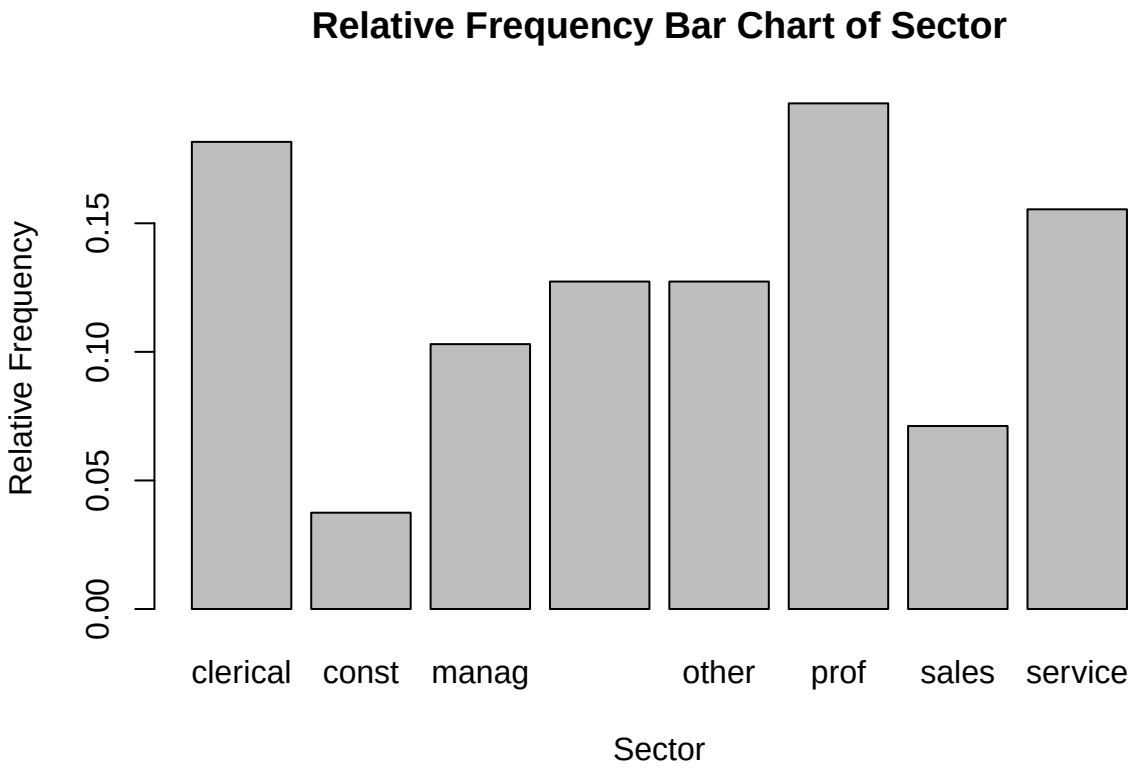
```
barchart(cps$sector, #This specifies the dataset and the variable
         horizontal = FALSE, #Turn the bars so they are vertical
         main = "Frequency Bar Chart of Sector", #Give your chart a title
         xlab = "Sector", #Label the x axis
         ylab = "Frequency", #Label the y axis
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



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

We could also choose to display the data as a proportion in a relative frequency bar chart. To find the relative frequency divide the count in each sector by the sample size. This is the sample proportion.

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



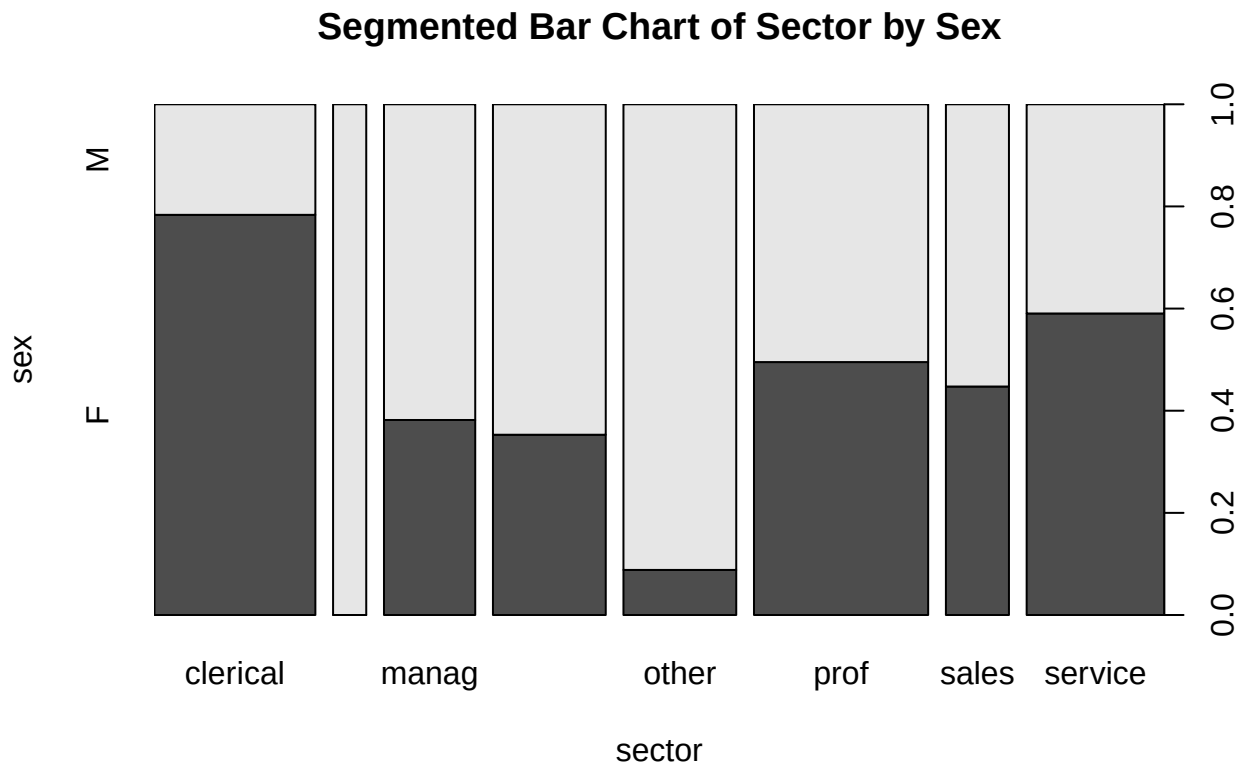
4. How does this plot differ from the plot above?

Displaying two categorical variables

To visually display two categorical variables we will use a segmented bar chart. In a segmented bar chart each bar sums to 100% or a proportion of 1. Typically the explanatory variable will be plotted on the y-axis and the response variable on the x-axis.

To see the differences in proportion of each sector between males and females we would create a segmented bar chart of sector segmented by sex. In this plot we are comparing the variable sex (explanatory variable) for the different sectors (response variable).

```
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
     )
```



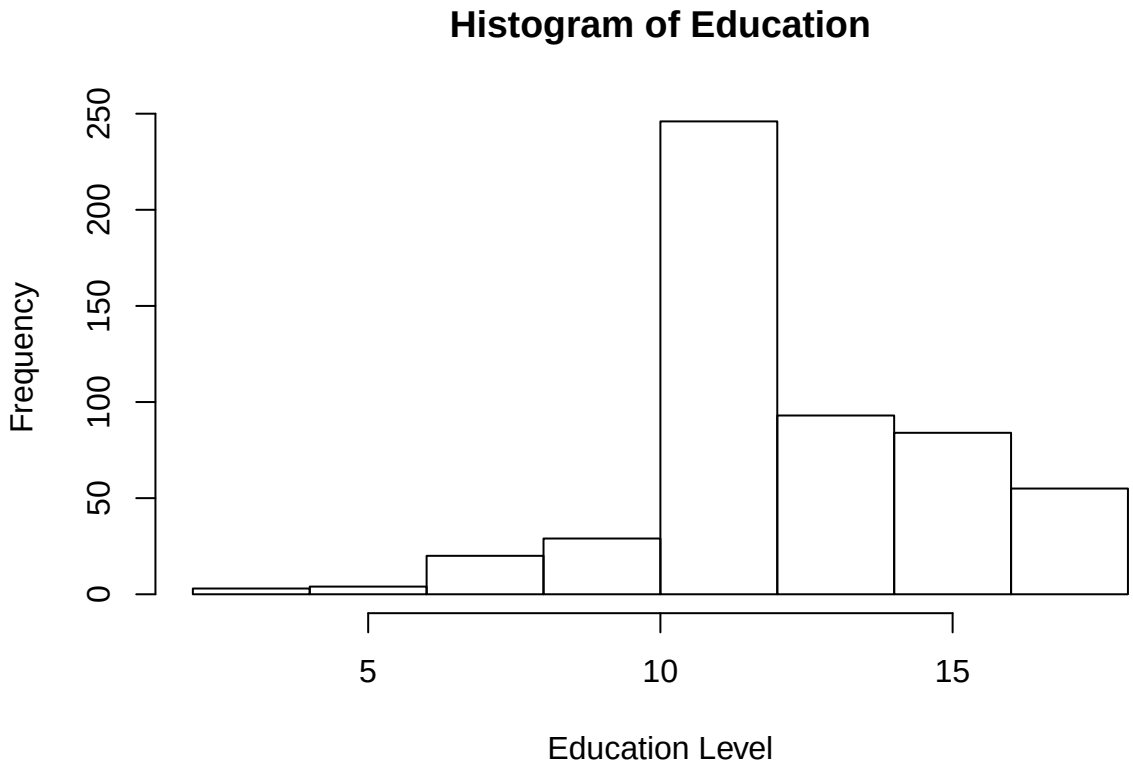
5. Using the segmented bar chart, which sector has about the same proportion of males and females?

6. Which sector has the highest proportion of females?

Displaying a single quantitative variable

To plot quantitative variables, we can use a dotplot, histogram or boxplot. To create a histogram the variable is broken into bins on a set width. Each bin plots the frequency of each variable for a certain bin. The following creates a histogram of the variable education. Notice that the bin width is 2 years. For example the first bin consists of the number of people in the data set of 0 to 2 years of education. It is important to note that someone with 2 years of education will fall into the bin for 2 - 4 years.

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



7. Which eduation levels have the highest frequency?

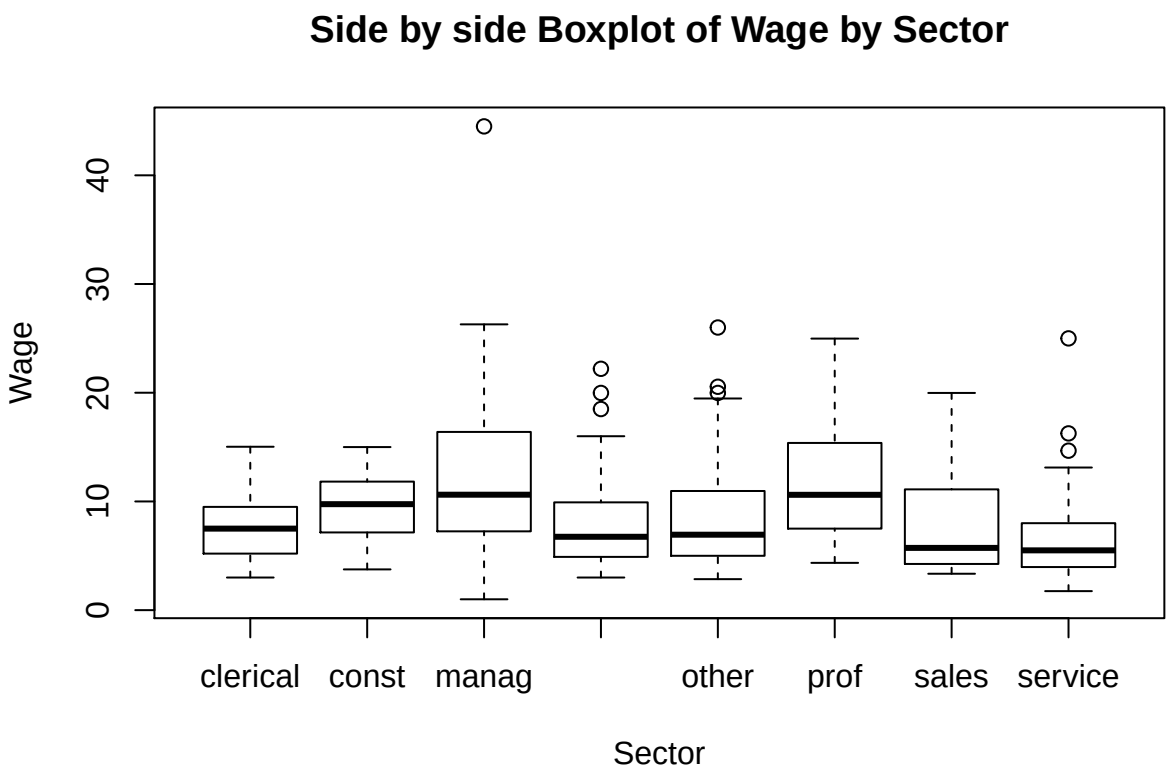
Displaying a Single Categorical and Single Quantitative Variable

Side by side plots are created from a single categorical and single quantitative variable. We can create side by side boxplots, side by side histograms, and side by side dotplots.

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

The boxplot of wage by sector is plotted using the code below. This plot helps to compare the wages for different sectors.

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



When comparing distributions we will look at four characteristics:

- Shape: A distribution that has approximately the same trailing off in both tails is considered a symmetric distribution. One in which the data trails off to the right is a right-skewed distribution. If the data trails off to the left it is considered a left-skewed distribution.
- Center: There are two measures of center used for quantitative data: the mean or the average and the median. The mean is found by adding up all the values in the dataset and dividing the sum by the sample size. The median is the 50th percentile of the ordered dataset.
- Spread: The spread is also referred to as variability. Again there are two measures of spread used to describe the data: the standard deviation and the Inter Quartile Range or IQR. The standard deviation measures the average distance of each observation from the mean. The IQR is found by subracting the value of the first quartile from the third quartile and measures the middle 50% of the data. $IQR = Q_3 - Q_1$
- Outliers: Outliers are values less than $Q_1 - 1.5 * IQR$ and greater than $Q_3 + 1.5 * IQR$. On the box plot they are displayed as a single dot beyond the whisker.

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

. . . Fill in the following table. . .