Discussion Section: Visualizing Data

Sam Frederick
sdf2128@columbia.edu

Administrative Notes

- Section slides will be posted on my Github
  - https://www.github.com/SamuelFrederick
- Office Hours Friday from 8:30 a.m. to 10:30 a.m.
  - Cafe on the 6th Floor of IAB

Example: Cooperative Election Study (CES)

- Large survey conducted annually
- Intended to be representative of congressional districts
- Scholars from different universities add questions
- We'll look at the 2020 CES with 61,000 respondents

Example: CES

- Download “ces.csv” from Courseworks > section > section_oct_5_6
- Move “ces.csv” to your course folder
- Set your working directory to your course folder:

```
setwd("~/Desktop/TA - Research Design Data Analysis/")
```

Example: CES

- Read in “ces.csv” and assign it to an object:

```
survey <- read.csv("ces.csv")
```

- Take a look at the data:

```
head(survey[, 1:5])
```

```
##   birthyr gender         educ  race    region
## 1    1966   Male       2 Year White Northeast
## 2    1955 Female    Post-Grad White     South
## 3    1946 Female       4 Year White   Midwest
## 4    1962 Female       4 Year White Northeast
## 5    1967   Male       4 Year White   Midwest
## 6    1961   Male Some College White   Midwest
```

```
dim(survey)
```

```
## [1] 61000    10
```

Example: CES

- Take the difference between the year 2020 and the variable ‘birthyr’
- Assign this difference to a new variable called ‘age’

Example: CES

- Take the difference between the year 2020 and the variable ‘birthyr’
- Assign this difference to a new variable called ‘age’

```
survey$age <- 2020 - survey$birthyr
```

Descriptive Statistics

- Central Tendency:
  - Median: median()
  - Mean: mean()

Descriptive Statistics

- Central Tendency:
    - Median: median()
    - Mean: mean()
- Spread:
    - Variance: var()
    - Standard deviation: sd()
    - IQR: IQR()
    - Range: range()
- Overall:
    - summary()

Example: CES

- Find the median and mean of the ‘age’ variable we created
- Find the standard deviation of the ‘age’ variable
- What is the maximum age in the sample?
- What is the minimum age in the sample?

Example: CES

- Find the median and mean of the ‘age’ variable we created

```
median(survey$age)
```

```
## [1] 49
```

```
mean(survey$age)
```

```
## [1] 48.38757
```

- Find the standard deviation of the ‘age’ variable

```
sd(survey$age)
```

```
## [1] 17.65902
```

Example: CES

- What is the maximum age in the sample?

```
max(survey$age)
```

```
## [1] 95
```

- What is the minimum age in the sample?

```
min(survey$age)
```

```
## [1] 18
```

```
summary(survey$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   18.00   33.00   49.00   48.39   63.00   95.00
```

Example: CES

- Choose a categorical variable from the dataset
- How would you summarize this variable?

Example: CES

- Choose a categorical variable from the dataset
- How would you summarize this variable?

```
table(survey$region)
```

```
## 
##    Midwest Northeast     South      West 
##      13667     11456     23493     12384
```

```
prop.table(table(survey$region))
```

```
## 
##    Midwest Northeast     South      West 
## 0.2240492 0.1878033 0.3851311 0.2030164
```

Visualizing Data

Data Type	Number of Variables	Plot Type	R Function
Categorical Data	1	Barplot	barplot()
Numeric Data	1	Histogram	hist()
Mixed Data	2+	Boxplot	boxplot()

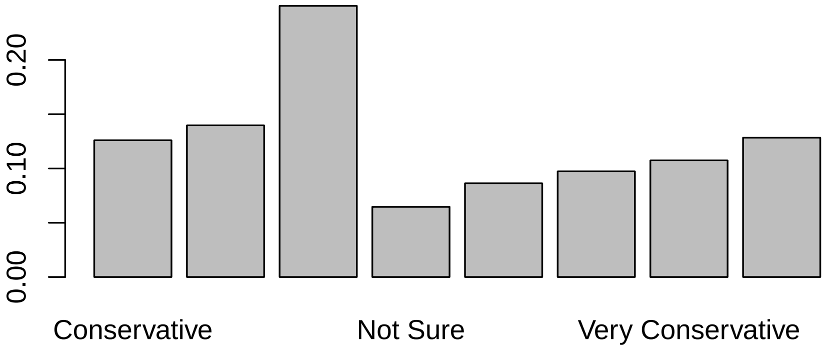
Barplots

- Plot distribution of factor/categorical variables
- Important considerations:
  - Ordering of factors
  - Labeling of factor levels in plot
- Helpful functions:
  - table()
  - prop.table()
  - tapply()

Example: CES

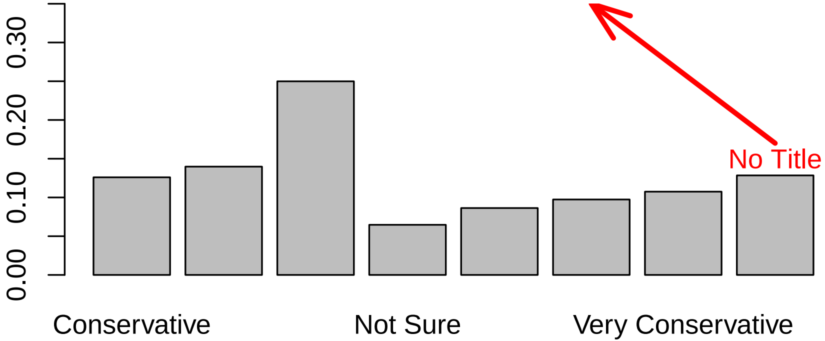
- What's wrong with this barplot?

```
barplot(prop.table(table(survey$ideology)))
```



Example: CES

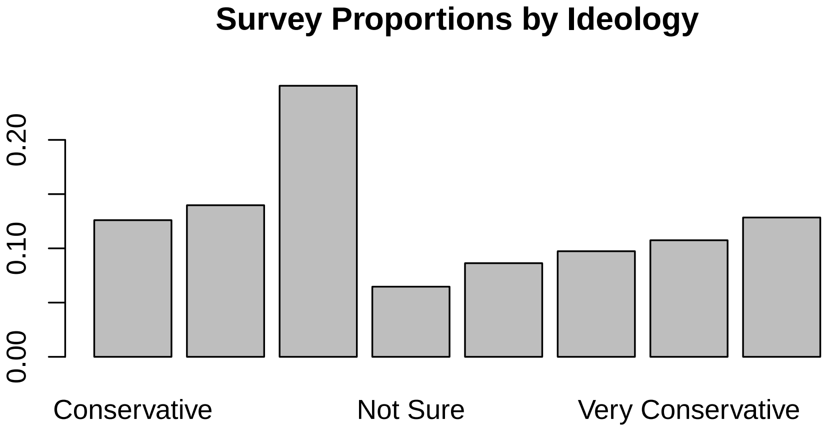
- What's wrong with this barplot?



No Title

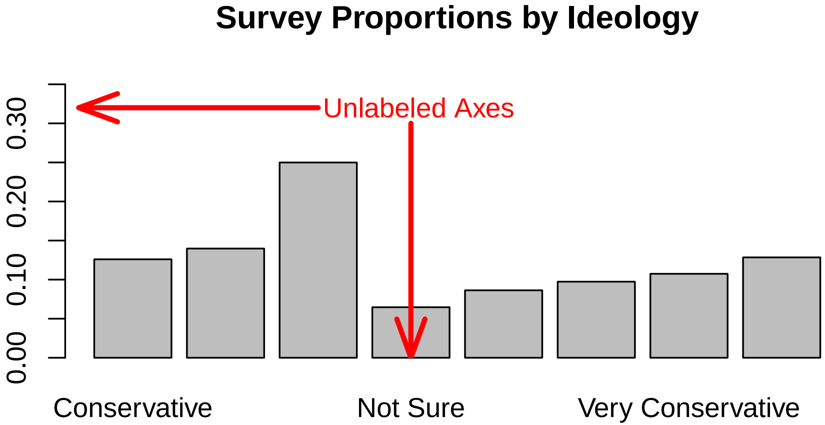
Example: Adding a Title

```
barplot(prop.table(table(survey$ideology)),
    main = "Survey Proportions by Ideology")
```



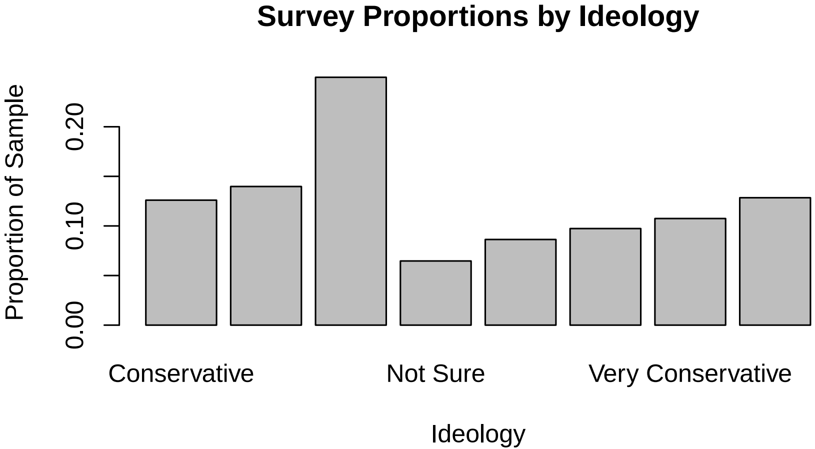
Example: CES

- What's wrong with this barplot?



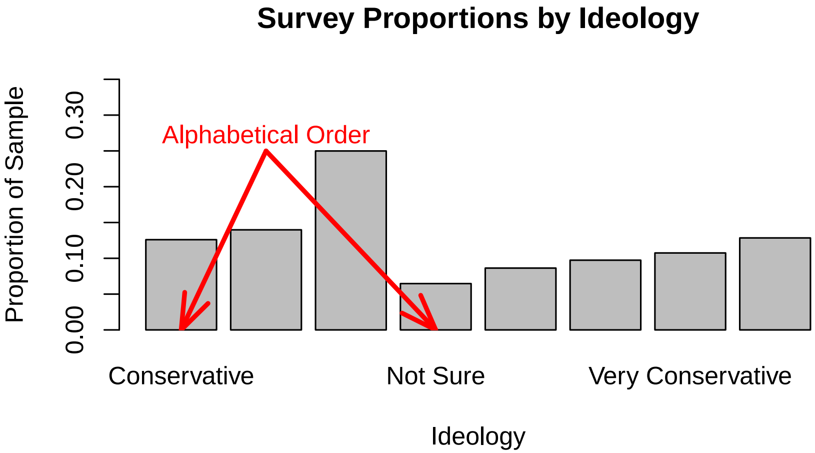
Example: Labeling Axes

```
barplot(prop.table(table(survey$ideology)),
    main = "Survey Proportions by Ideology",
    xlab = "Ideology", ylab = "Proportion of Sample")
```



Example: CES

- What's wrong with this barplot?



Example: Ordering Factors

```
unique(survey$ideology)
```

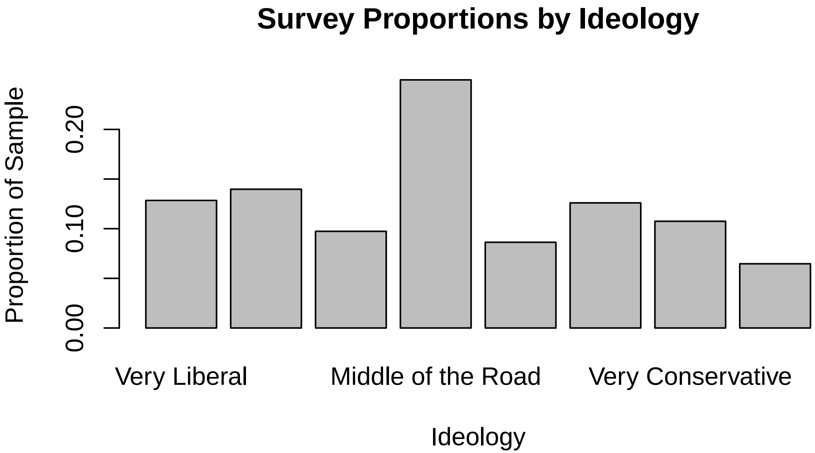
Example: Ordering Factors

```r
unique(survey$ideology)
```

```r
survey$ideology <- factor(survey$ideology,
    levels = c("Very Liberal", "Liberal",
        "Somewhat Liberal", "Middle of the Road",
        "Somewhat Conservative", "Conservative",
        "Very Conservative", "Not Sure"))
```

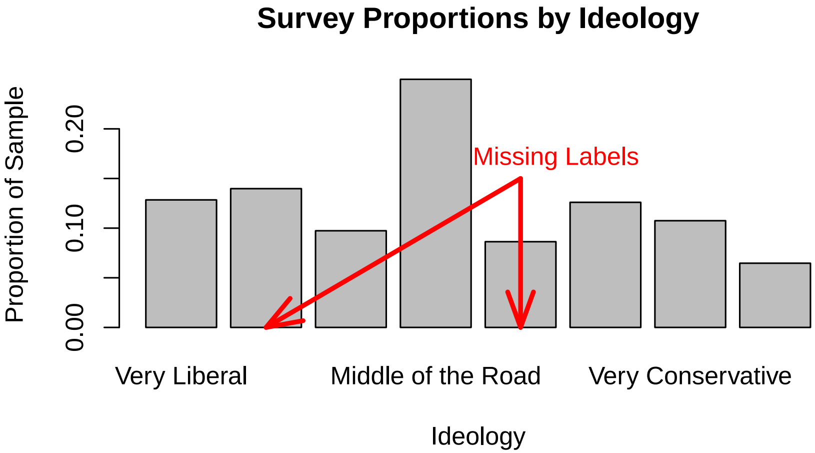
Example: Ordering Factors

```
barplot(prop.table(table(survey$ideology)),
    main = "Survey Proportions by Ideology",
    xlab = "Ideology", ylab = "Proportion of Sample")
```



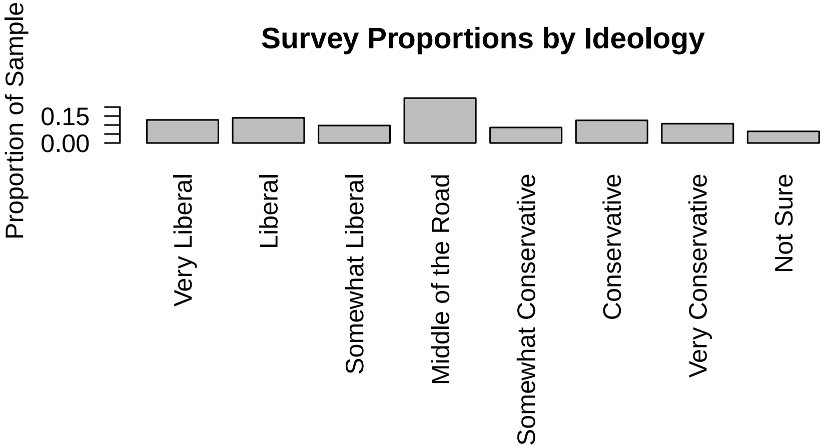
Example: CES

- What's wrong with this barplot?



Example: Dealing with Long Factor Names

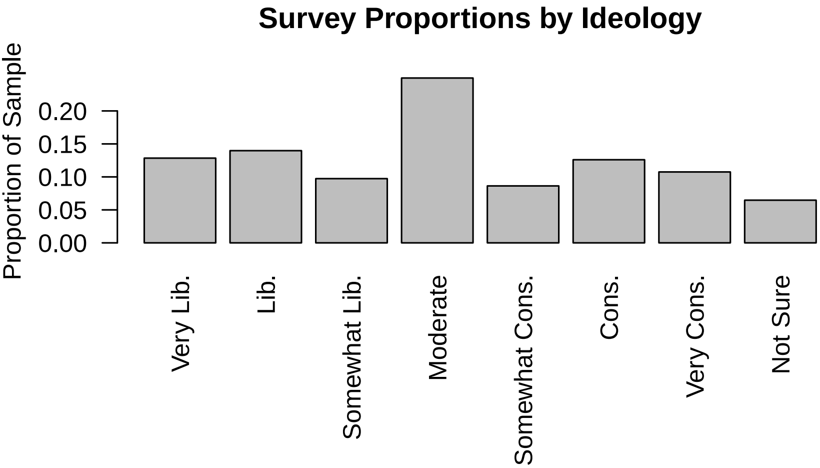
```
par(mar = c(12, 4, 4, 2))
barplot(prop.table(table(survey$ideology)),
    main = "Survey Proportions by Ideology",
    ylab = "Proportion of Sample", las = 2)
```



Example: Dealing with Long Factor Names

```r
survey$ideology <- factor(survey$ideology,
    levels = c("Very Liberal", "Liberal",
        "Somewhat Liberal", "Middle of the Road",
        "Somewhat Conservative", "Conservative",
        "Very Conservative", "Not Sure"),
    labels = c("Very Lib.", "Lib.", "Somewhat Lib.",
        "Moderate", "Somewhat Cons.", "Cons.",
        "Very Cons.", "Not Sure"))
par(mar = c(8, 4, 4, 2))
barplot(prop.table(table(survey$ideology)),
    main = "Survey Proportions by Ideology",
    ylab = "Proportion of Sample", las = 2)
```

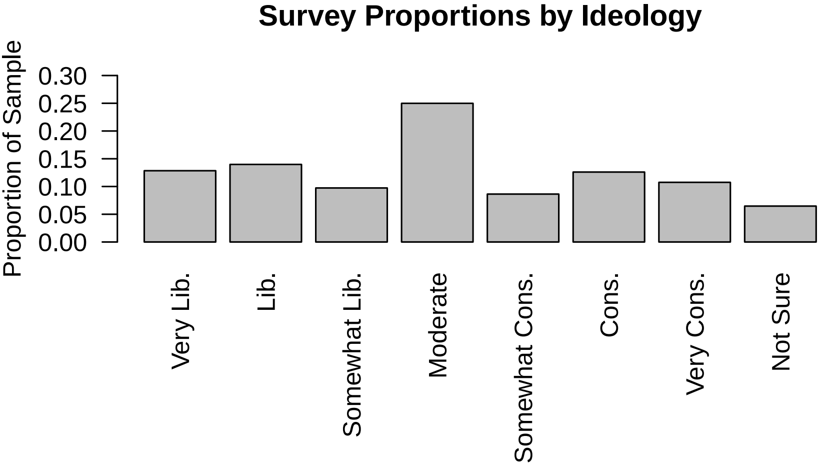
Example: Dealing with Long Factor Names



Example: Plot Limits

```r
par(mar = c(8, 4, 4, 2))
barplot(prop.table(table(survey$ideology)),
    main = "Survey Proportions by Ideology",
    ylab = "Proportion of Sample", las = 2,
    ylim = c(0, 0.3))
```

Example: Plot Limits

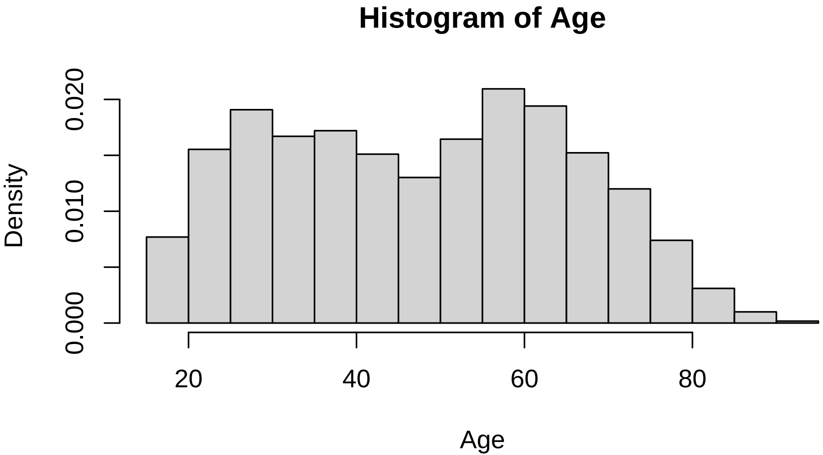


Histograms

- Plot distribution of numeric variables
- $height = \frac{\text{Proportion of observations within an interval}}{\text{Interval width}}$
- $height * Interval\ width =$
  $Proportion\ of\ observations\ within\ an\ interval$

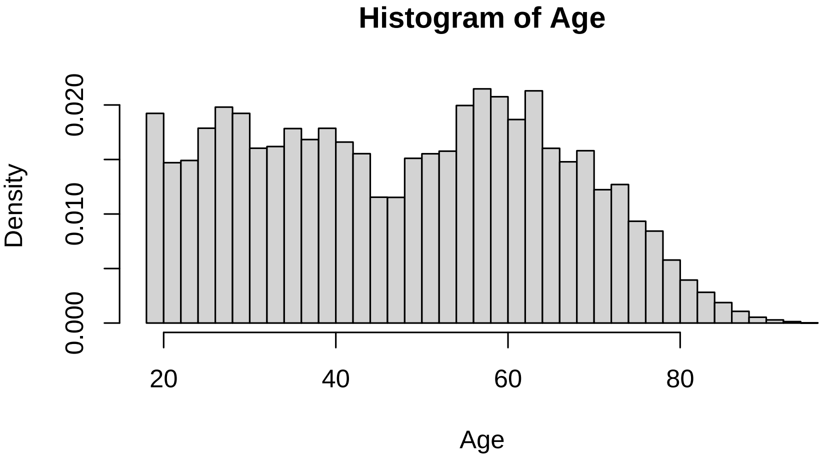
Example: Histogram of Age

```r
hist(survey$age, freq = FALSE, main = "Histogram of Age",
    xlab = "Age")
```



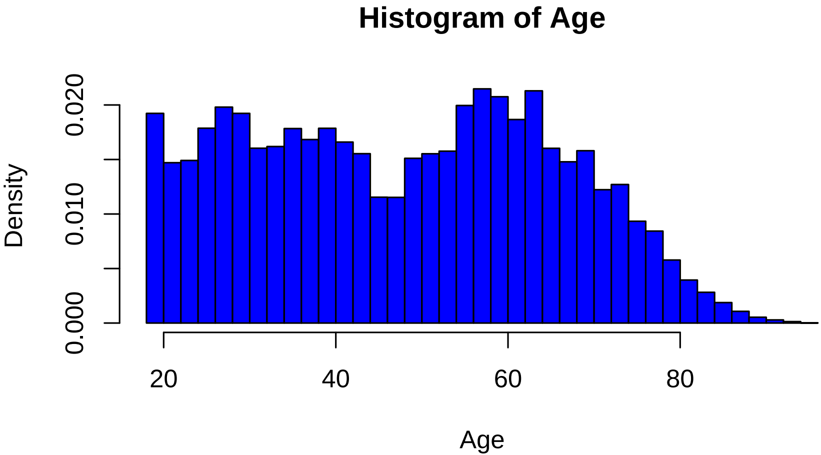
Example: Histogram of Age

```
hist(survey$age, freq = FALSE, main = "Histogram of Age",
    xlab = "Age", breaks = 30)
```



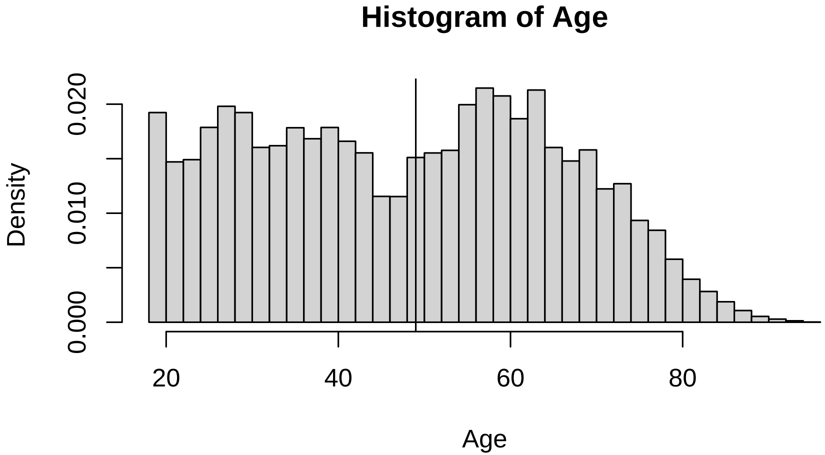
Example: Histogram of Age

```r
hist(survey$age, freq = FALSE, main = "Histogram of Age",
     xlab = "Age", breaks = 30, col = "blue")
```



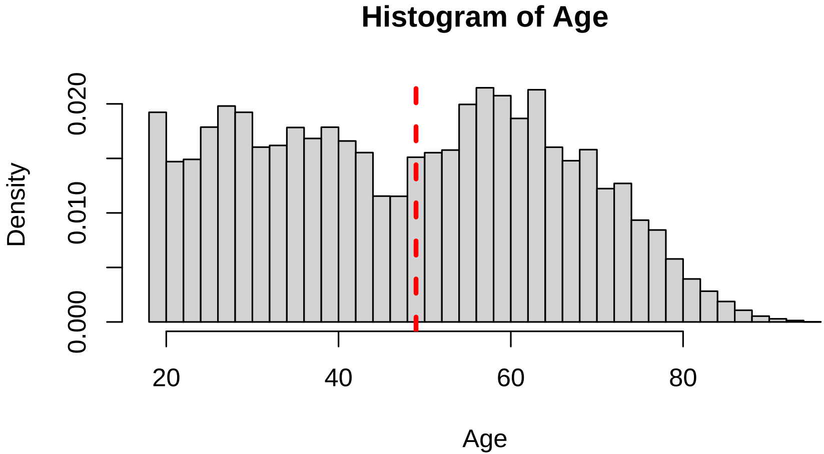
Example: Histogram of Age

```
hist(survey$age, freq = FALSE, main = "Histogram of Age",
    xlab = "Age", breaks = 30)
abline(v = median(survey$age))
```



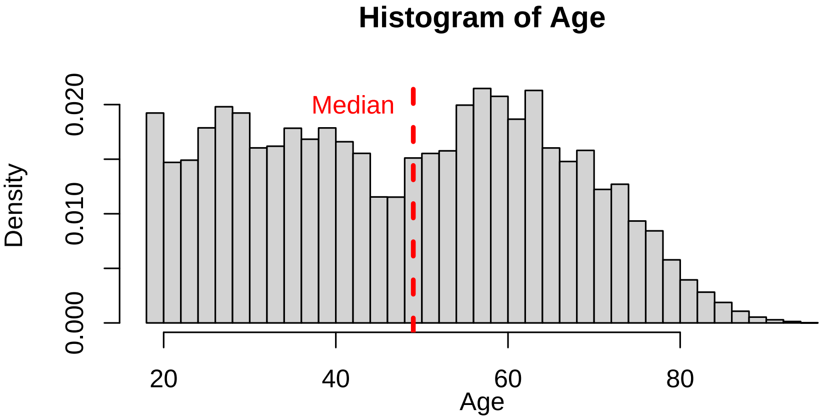
Example: Histogram of Age

```r
hist(survey$age, freq = FALSE, main = "Histogram of Age",
    xlab = "Age", breaks = 30)
abline(v = median(survey$age), col = "red",
    lty = "dashed", lwd = 3)
```



Example: Histogram of Age

```r
hist(survey$age, freq = F, main = "Histogram of Age",
    xlab = "Age", breaks = 30)
abline(v = median(survey$age), col = "red",
    lty = "dashed", lwd = 3)
text(x = 42, y = 0.02, "Median", col = "red")
```

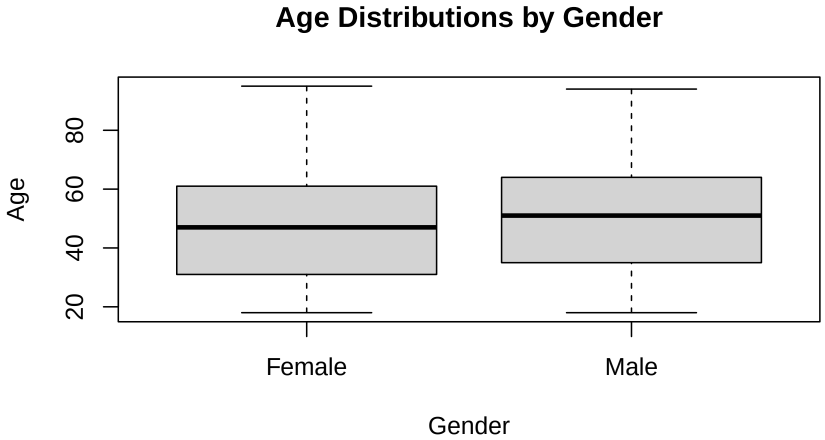


Boxplots

- Helpful for comparing the distributions of numeric variables
- Shows Interquartile Range, Median, and Outliers

Example: Boxplot

```
boxplot(age ~ gender, data = survey, xlab = "Gender",
    ylab = "Age", main = "Age Distributions by Gender")
```



Random Sampling

- Randomization ensures groups similar on average
    - Random sampling $\implies$ representative samples
- Still biases:
    - Social Desirability Bias
    - Non-response
    - Item non-response

Sampling in R

- Random tasks in R will vary each time we run them
- To ensure replicability, we set a seed

```
set.seed(123)
```

- To take a random sample, we use the sample() function

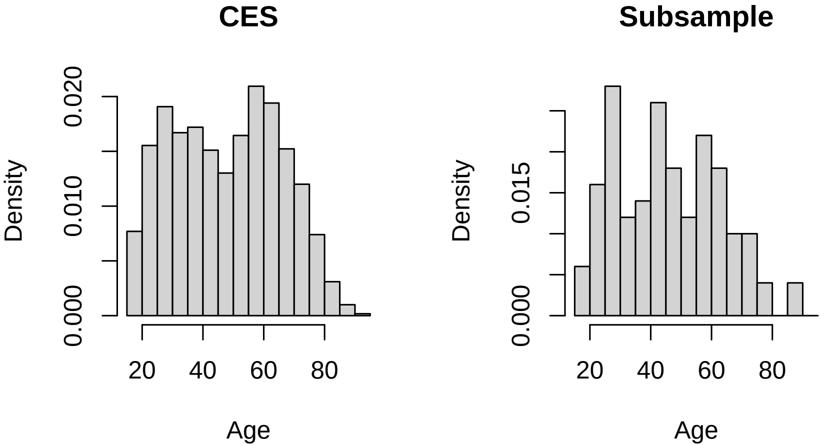
```
sample(10, size = 3, replace = FALSE)
```

```
## [1]  3 10  2
```

Example

```r
set.seed(123)
survey_sample <- survey[sample(nrow(survey),
    size = 100, replace = FALSE), ]
par(mfrow = c(1, 2))
hist_age <- hist(survey$age, freq = FALSE,
    xlab = "Age", main = "CES")
samp_hist_age <- hist(survey_sample$age,
    breaks = hist_age$breaks, freq = FALSE,
    xlab = "Age", main = "Subsample")
```

Example

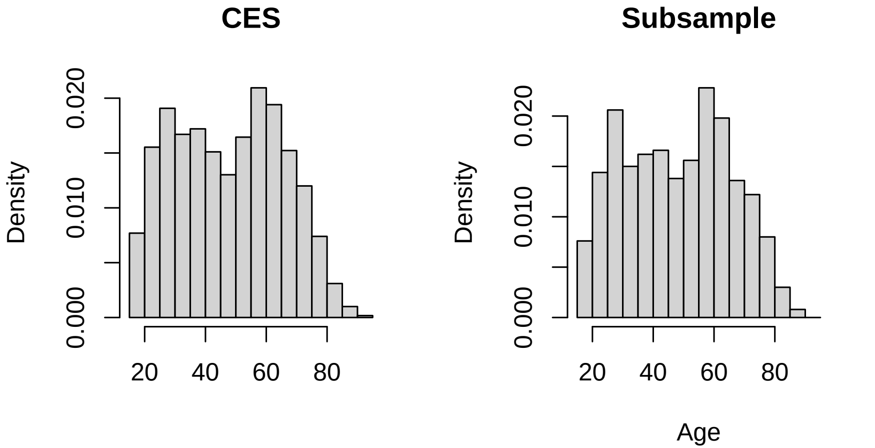

CES
Density
0.000
0.010
0.020
20
40
60
80
Age
Subsample
Density
0.000
0.015
20
40
60
80
Age

Example

```r
set.seed(123)
survey_sample <- survey[sample(nrow(survey),
    size = 1000, replace = FALSE), ]
```

```r
par(mfrow = c(1, 2))
hist_age <- hist(survey$age, freq = FALSE,
    xlab = "Age", main = "CES")
samp_hist_age <- hist(survey_sample$age,
    breaks = hist_age$breaks, freq = FALSE,
    xlab = "Age", main = "Subsample")
```

Example


CES
Density
0.000
0.010
0.020
20
40
60
80
Subsample
Density
0.000
0.010
0.020
20
40
60
80
Age