# Histograms and Plots

Heather Dye

November 4, 2022

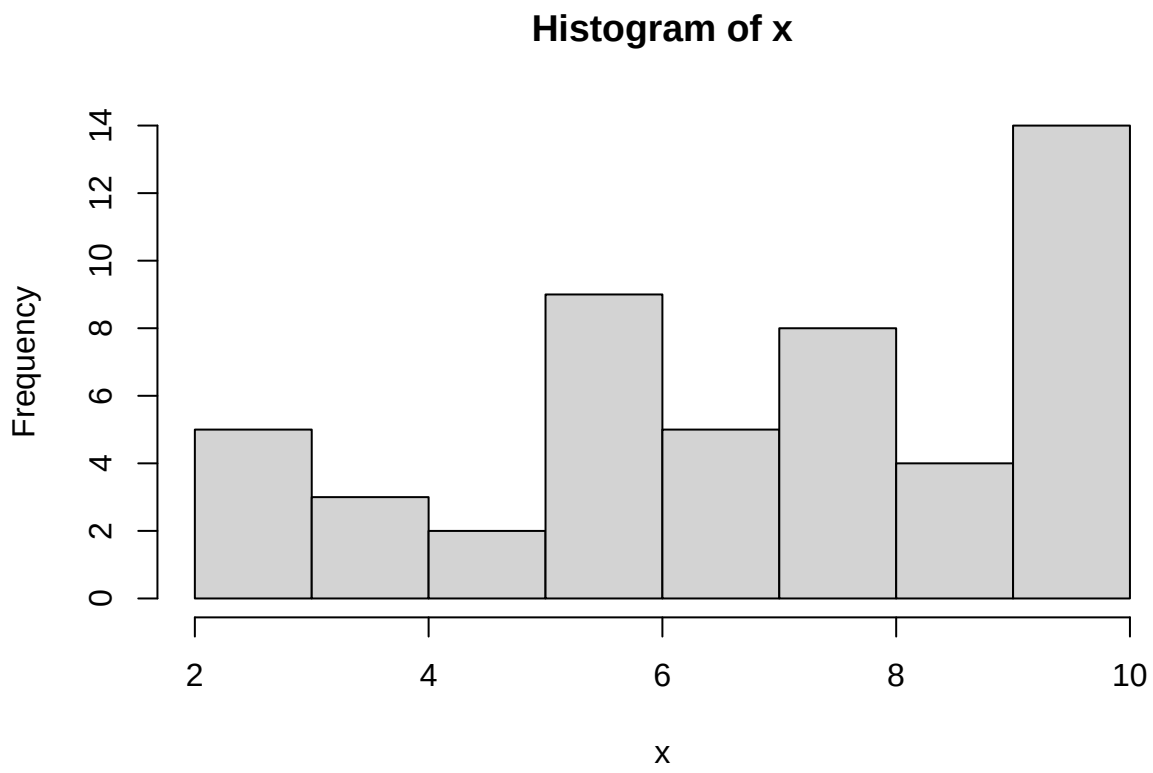
## Descriptive Statistics - Histograms and Plots

Given a data set, we want to determine the distribution that the data comes from and the defining parameters.

We will learn how to construct histograms. We'll start by examining some basic distributions and learning descriptive terms.

**Example 1** What distribution do you think this is from?

```
x=runif(50, min=2, max =10)
hist(x)
```




This is from a uniform distribution with a minimum of 2 and a maximum of 10. The structure of the histogram creates the impression that the maximum value of the data set is 10 and the minimum value is 2. But, they are actually:

```
max(x)
```

```
## [1] 9.911134
```

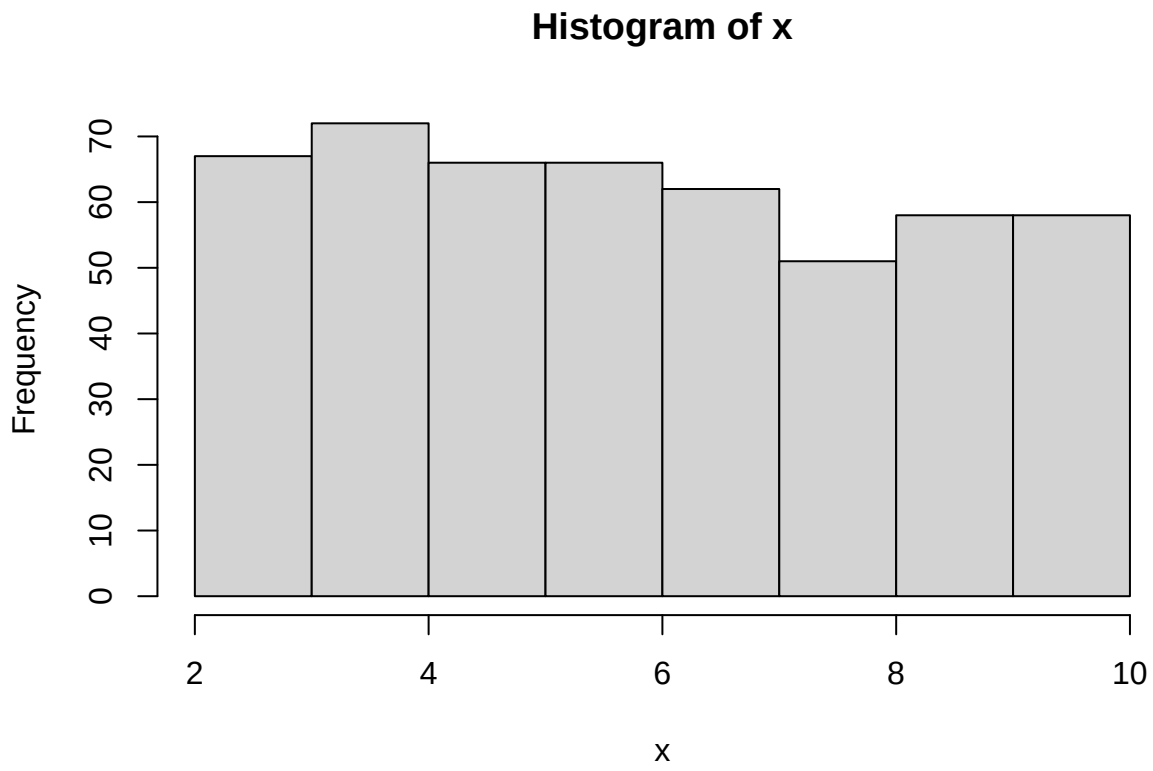
```
min(x)
```

```
## [1] 2.031587
```

The data values get close to these extremes, but without prior knowledge it is difficult to guess the values exactly. Based on the data set, the distribution could be from 2.03 to 9.99.

Now, what distribution do you think this is from?

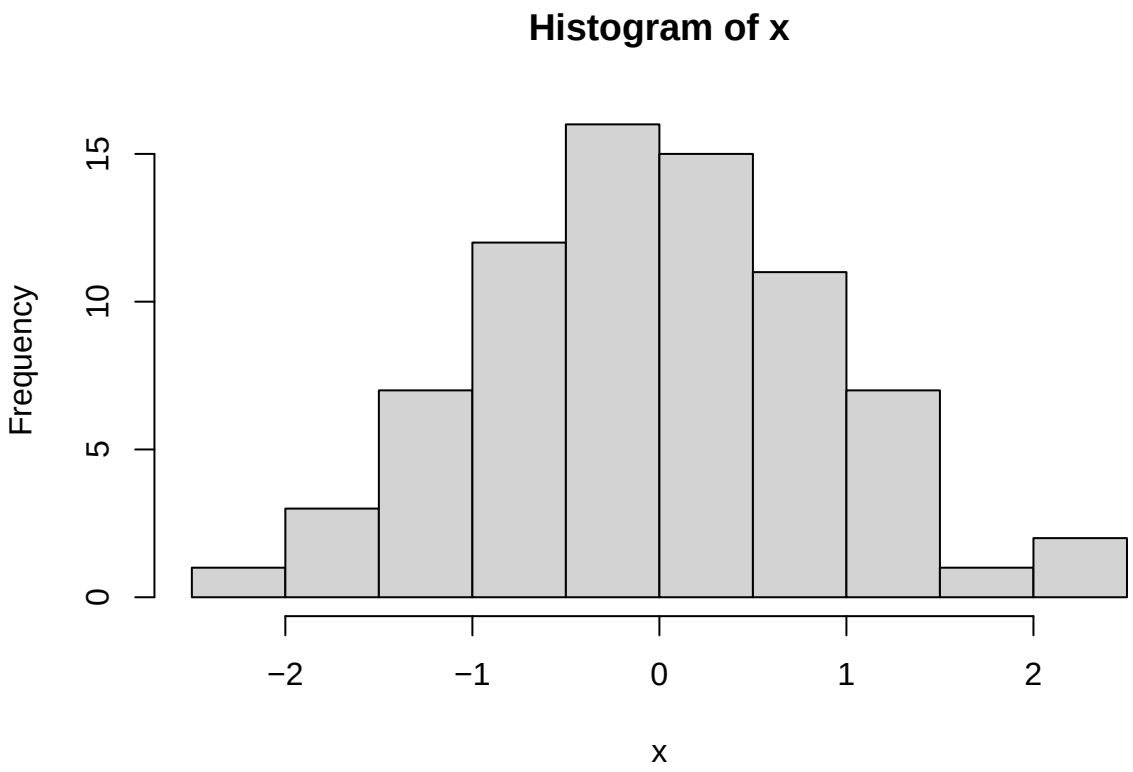
```
x=runif(500, min=2, max =10)
hist(x)
```




This is the same distribution, but with a larger data set. You can see how the additional data values *fill in* the distribution, creating a more accurate picture of the shape of the population.

**Example 2** What distribution do you think this data is from?

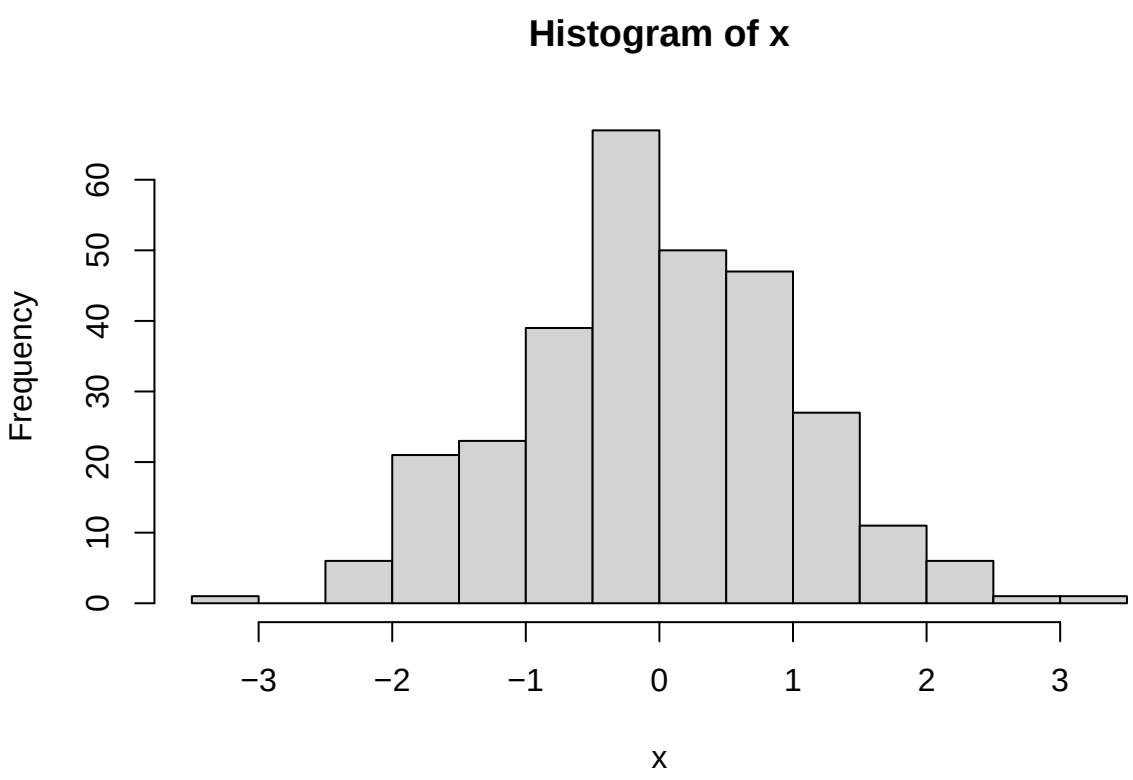
```
x = rnorm(75)
hist(x)
```




This data set is drawn from a normal distribution with mean zero and variance one. The distribution is symmetric and bell shaped (the bell curve).

Take 2:

```
x = rnorm(300)
hist(x)
```




Again, we see that larger data sets "fill in" the shape of the distribution.

**Example 3: Steps to computing a histogram by hand.**

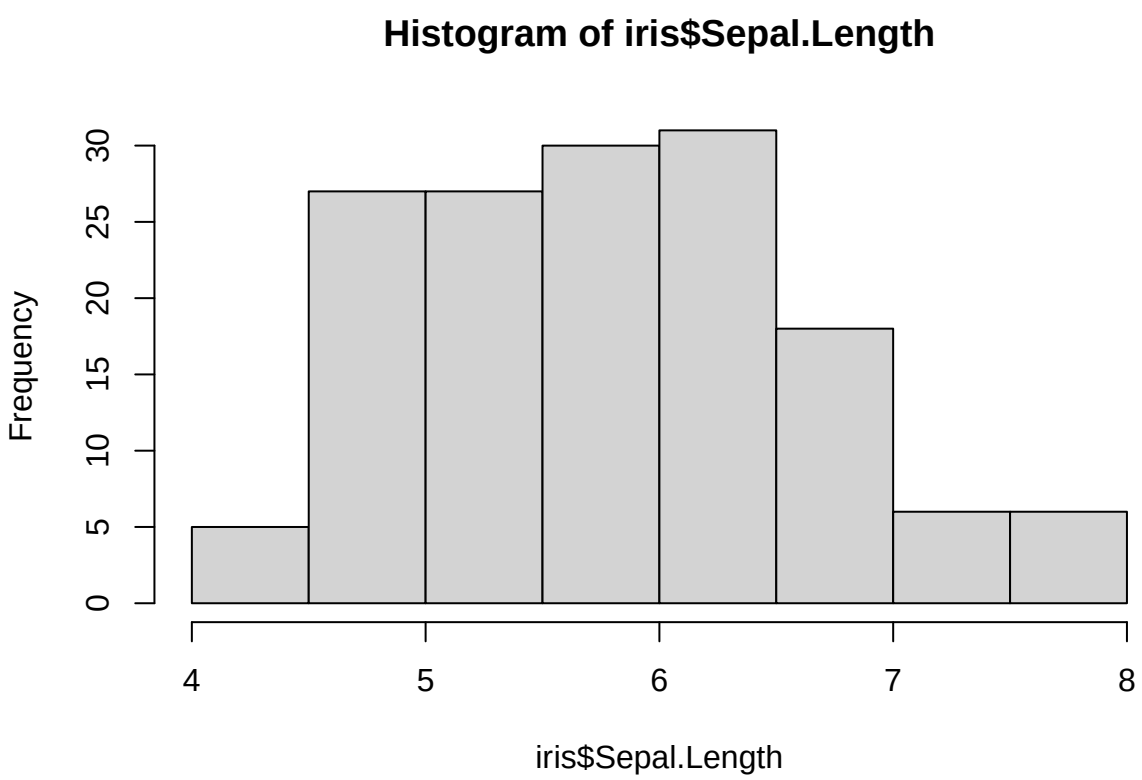
- Compute the range = max - min
- Select $k$, the number of intervals. In general, $k$ is between 5 and 20.
- Estimate the size of an interval: range/$k$. Round to a reasonable number.
- Construct the intervals: $[c_0, c1), [c_1, c2) \ldots [c_{k-1}, c_k)$. The endpoints are called class boundaries.
- Choose frequency or relative frequency. Graph.

**Common Terms describing the shape of data:**

- Skew (to the left, to the right)
- Normal
- Uniform
- Bimodal

**Example 4** Construct a histogram of the Sepal.Length variable in the data set iris. Note: This data set is built into R datasets, but you can also download it from the UCI Machine Learning Repository.

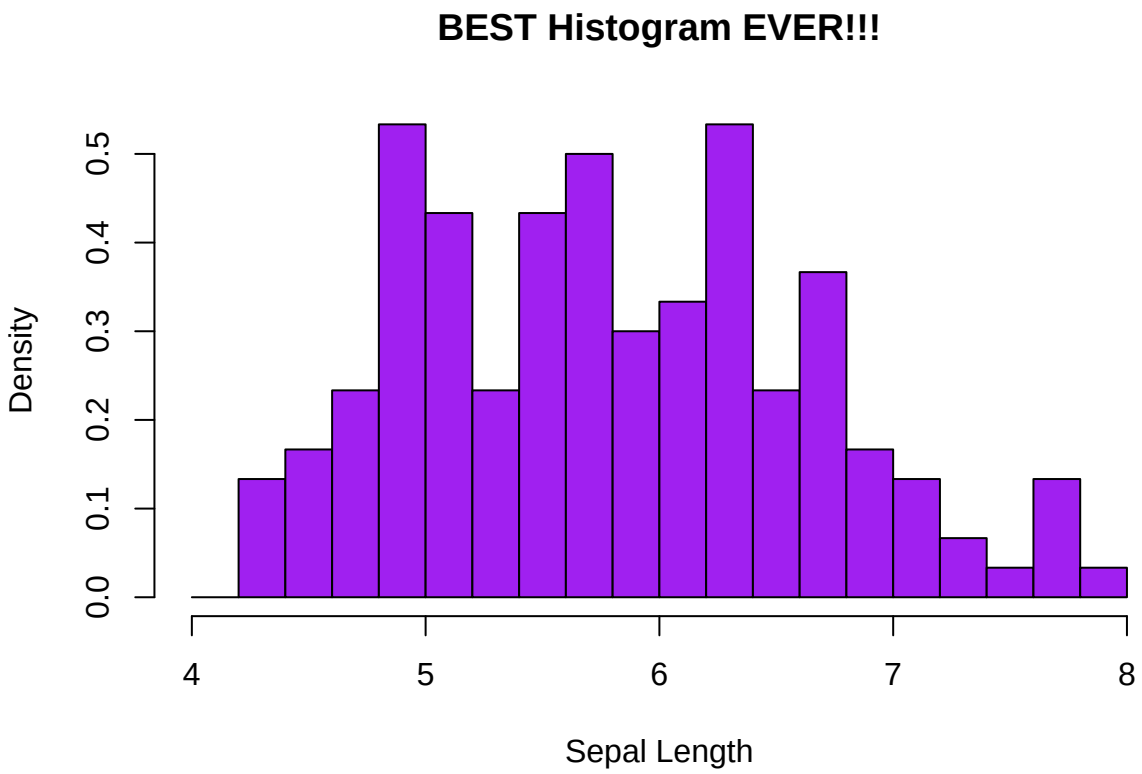
```
hist(iris$Sepal.Length)
```




We can customize the plot. Note that the maximum is 7.9 and the minimum value of the data set is 4.3. Then the range of the data set is 3.6. I choose $k = 20$ and the estimated length of an interval is 0.18. I round this value to 0.20.

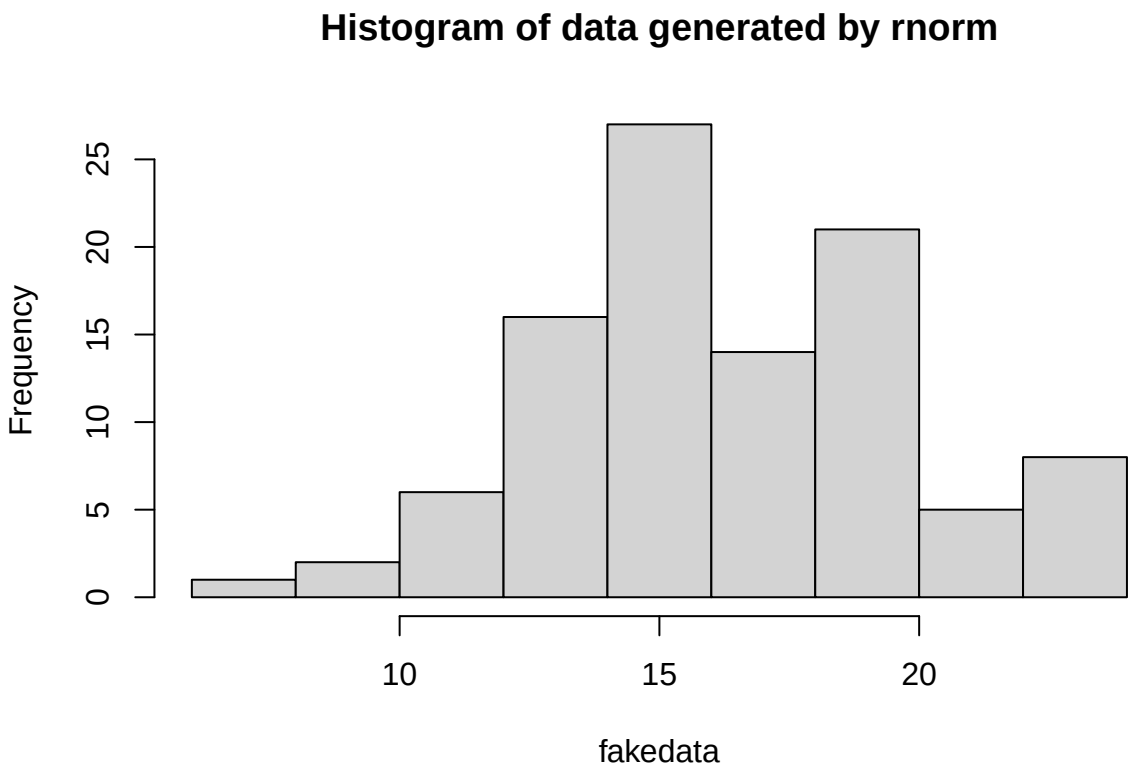
```
vec<-seq(from=4,to=8, by=0.2)

hist(iris$Sepal.Length, breaks=vec, freq=FALSE, col="purple", main="BEST Histogram EVER!!!", xlab="Sepal
```




**Example 5: Fake normal data** We construct some fake data using the rnorm command. Then, we construct a histogram. In this case, we expect to see a bell curve centered around 17 and most of the data should be in the interval $(5, 29)$.

```
fakedata<-rnorm(100, mean = 17, sd = 4)
hist(fakedata, main="Histogram of data generated by rnorm")
```




**Example 6: Histogram with CSV File** This example uses a cereal data set from Chris Crawford. The data is available at:

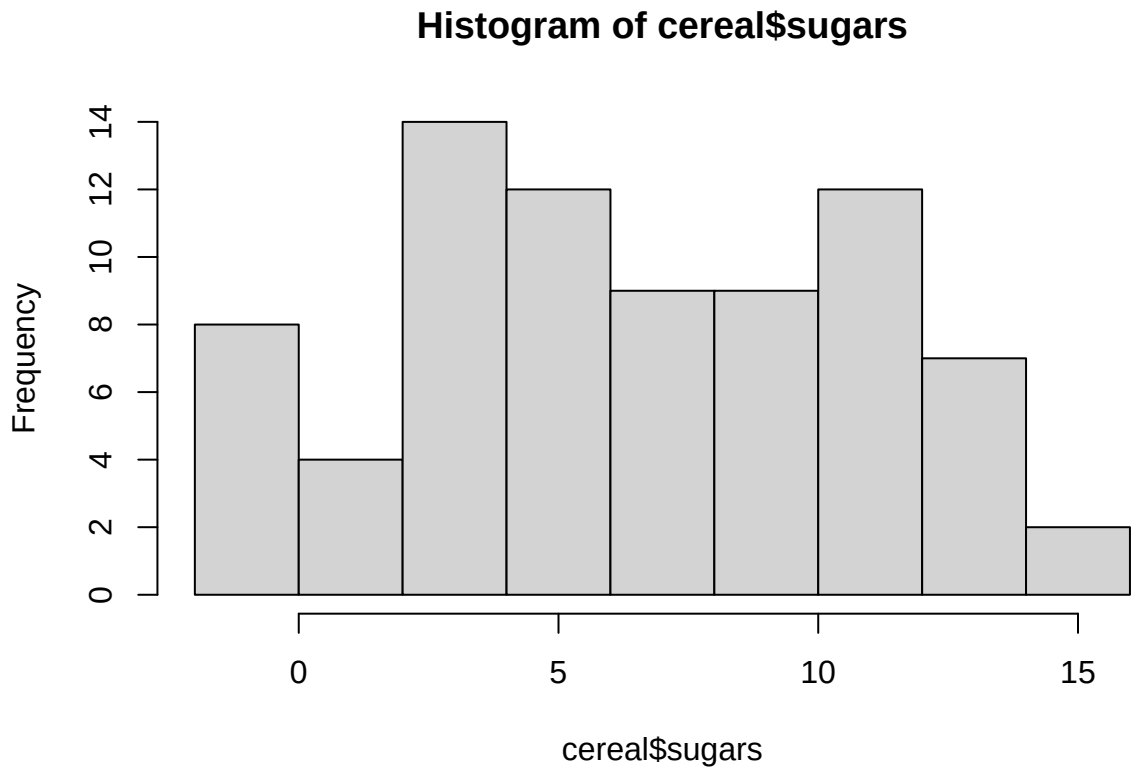
https://www.kaggle.com/datasets/crawford/80-cereals

- Set your working directory
- Read in the csv file
- Identify the variable names
- Create a histogram of the sugar levels in the cereals.
- Then, identify the cereals with zero sugar.

```
cereal<-read.csv(file="cereal.csv", header = TRUE)
names(cereal)
```

```
##  [1] "name"     "mfr"      "type"     "calories" "protein"  "fat"
##  [7] "sodium"   "fiber"    "carbo"    "sugars"   "potass"   "vitamins"
## [13] "shelf"    "weight"   "cups"     "rating"
```

```
hist(cereal$sugars)
```




```
zerosugar<-cereal[cereal$sugars==0, ]
zerosugar$name
```

```
## [1] "All-Bran with Extra Fiber" "Cream of Wheat (Quick)"
## [3] "Puffed Rice"               "Puffed Wheat"
## [5] "Shredded Wheat"            "Shredded Wheat 'n'Bran"
## [7] "Shredded Wheat spoon size"
```