Assignment 1 Solutions

Abhishek Devarajan

2023-11-06

Part 1

1: Vectors

a) Create a vector named *heights* that contains the heights, in inches, of yourself and two students near you. Print the contents of this vector.

```
heights <- c(66, 70, 69)
heights
```

```
## [1] 66 70 69
```

b) Create a vector named *names* that contains the names of these people. Print the contents of this vector.

```
names <- c("Abhishek","Steven","Dan")
names
```

```
## [1] "Abhishek" "Steven"   "Dan"
```

c) Try typing cbind(heights, names). What did this command do? What class is this new object?

```
cbind(heights, names)
```

```
##      heights names
## [1,] "66"    "Abhishek"
## [2,] "70"    "Steven"
## [3,] "69"    "Dan"
```

The cbind() function took the vectors heights and names and combined them into a table. Each vector was given its own column in the table.

```
class(cbind(heights,names))
```

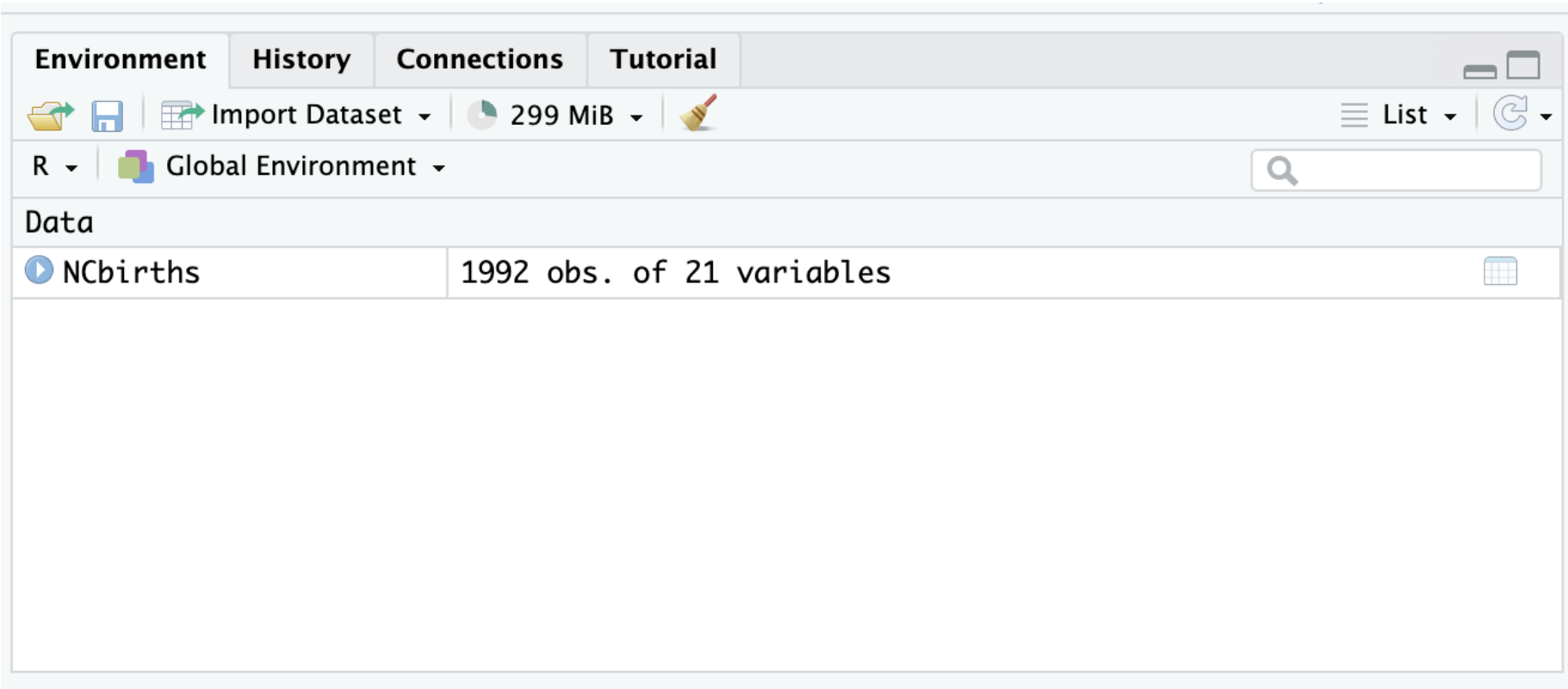
```
## [1] "matrix" "array"
```

Based on the class() function, it looks like cbind creates a matrix array.

2: Downloading Data

a) Download the data set births.csv from the course site and upload it into RStudio. Name the data frame NCbirths.

```
setwd("~/Documents/2ndYearPhD/Fall23-24/Stats10/")
NCbirths <- read.csv("births.csv")
```



NCbirths in the Environment tab

b) Demonstrate that you have been successful by typing head(NCbirths) and copying and pasting the output into your word processing document.

```
head(NCbirths)
```

```
##   Gender Premie weight Apgar1 Fage Mage Feduc Meduc TotPreg Visits   Marital
## 1   Male     No    124      8   31   25    13    14       1     13   Married
## 2 Female     No    177      8   36   26     9    12       2     11 Unmarried
## 3   Male     No    107      3   30   16    12     8       2     10 Unmarried
## 4 Female     No    144      6   33   37    12    14       2     12 Unmarried
## 5   Male     No    117      9   36   33    10    16       2     19   Married
## 6 Female     No     98      4   31   29    14    16       3     20   Married
##   Racemom Racedad Hispmom Hispdad Gained     Habit MomPriorCond BirthDef
## 1   White   White NotHisp NotHisp     40 NonSmoker         None     None
## 2   White   White Mexican Mexican     20 NonSmoker         None     None
## 3   White Unknown Mexican Unknown     70 NonSmoker At Least One     None
## 4   White   White NotHisp NotHisp     50 NonSmoker         None     None
## 5   White   Black NotHisp NotHisp     40 NonSmoker At Least One     None
## 6   White   White NotHisp NotHisp     21 NonSmoker         None     None
##      DelivComp BirthComp
## 1 At Least One      None
## 2 At Least One      None
## 3 At Least One      None
## 4 At Least One      None
## 5         None      None
## 6         None      None
```

3: Packing Loading

a) Install the maps package. Verify its installation by typing find.package("maps") and include the output in your answer.

```
install.packages("maps")
```

```
## Installing package into '/opt/homebrew/lib/R/4.3/site-library'
## (as 'lib' is unspecified)
```

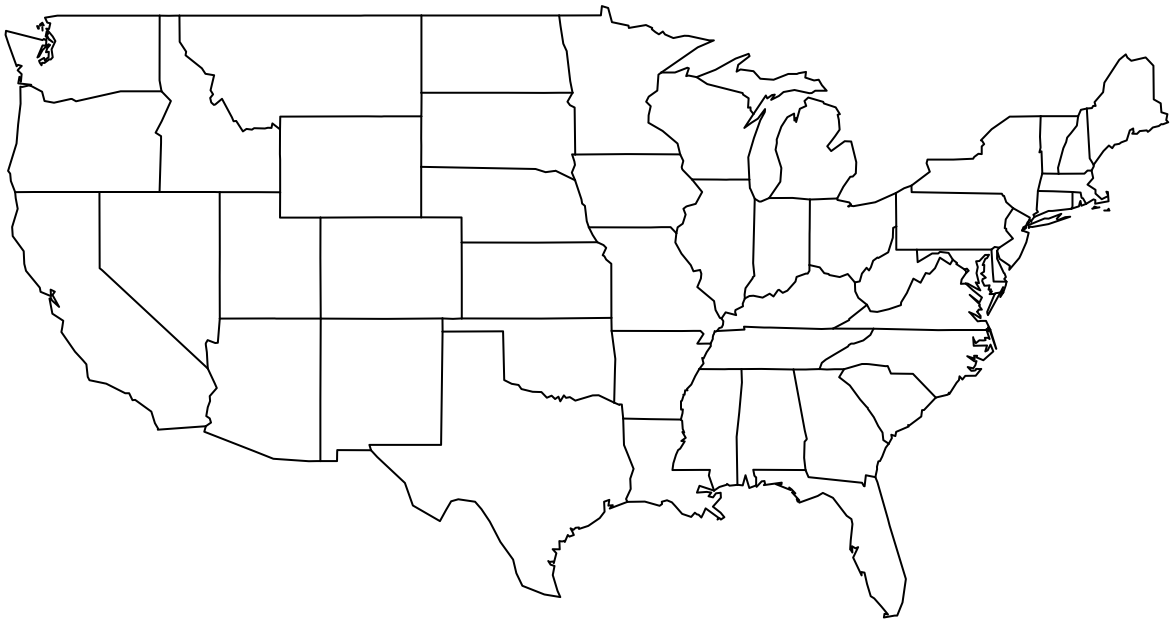
```
## Warning in install.packages("maps"): installation of package 'maps' had
## non-zero exit status
```

```
find.package("maps")
```

```
## [1] "/opt/homebrew/lib/R/4.3/site-library/maps"
```

b) Type library(maps) to load up the package. Type map("state") and include the plot output in your answer.

```
library(maps)
map("state")
```



4: Perform Vector Operations

a) Extract the weight variable as a vector from the data frame

```
weights <- NCbirths$weight

# Checking the output (optional)
weights[1:10]
```

```
##  [1] 124 177 107 144 117  98 147 138 104 123
```

b) What units do you think the weights are in?

Given that this is an American dataset and that we are measuring the weight of babies, the measurements are most likely given in ounces.

c) Create a new vector named weights_in_pounds which are the weights of the babies in pounds. You can look up conversion factors on the internet.

```
weights_in_pounds <- weights / 16
```

d) Demonstrate your success by typing weights_in_pounds[1:20] and including the output in your word processing document.

```
weights_in_pounds[1:20]
```

```
##  [1]  7.7500 11.0625  6.6875  9.0000  7.3125  6.1250  9.1875  8.6250  6.5000
## [10]  7.6875  9.5625  8.0625  7.4375  6.7500  6.6250  7.8125  7.1875  8.0000
## [19]  8.2500  5.1875
```

5: What is the mean weight of the babies in pounds?

```
mean(weights_in_pounds)
```

```
## [1] 7.2532
```

The average weight of the babies in pounds is around 7.25 pounds.

a) What percentage of the mothers in the sample smoke? *Hint: use the tally function with the format argument. Use the help screen for guidance*.

```
library(mosaic)
```

```
## Registered S3 method overwritten by 'mosaic':
##   method                           from
##   fortify.SpatialPolygonsDataFrame ggplot2
```

```
##
## The 'mosaic' package masks several functions from core packages in order to add
## additional features.  The original behavior of these functions should not be affected by this.
```

```
##
## Attaching package: 'mosaic'
```

```
## The following objects are masked from 'package:dplyr':
##
##     count, do, tally
```

```
## The following object is masked from 'package:Matrix':
##
##     mean
```

```
## The following object is masked from 'package:ggplot2':
##
##     stat
```

```
## The following objects are masked from 'package:stats':
##
##     binom.test, cor, cor.test, cov, fivenum, IQR, median, prop.test,
##     quantile, sd, t.test, var
```

```
## The following objects are masked from 'package:base':
##
##     max, mean, min, prod, range, sample, sum
```

```
tally(NCbirths$Habit, format = "percent")
```

```
## X
## NonSmoker    Smoker
##  90.61245   9.38755
```

```
#You can round the percentage however you want.
```

Based on the tally() function, about 9.4% of mothers in the NCbirths dataset smoke.

b) According to the Centers for Disease Control, approximately 21% of adult Americans are smokers. How far off is the percentage you found in 2 from the CDC's report?

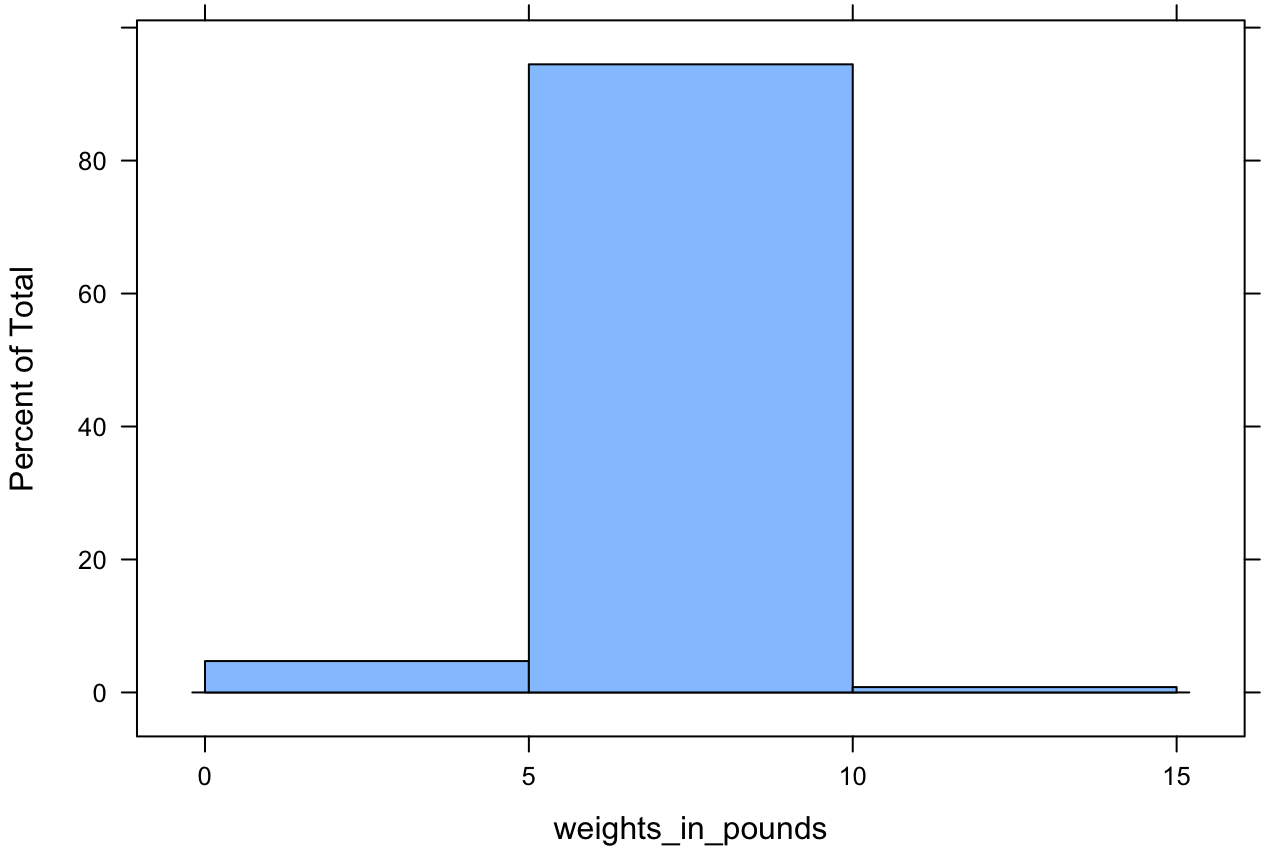
```
21 - 9.38755
```

```
## [1] 11.61245
```

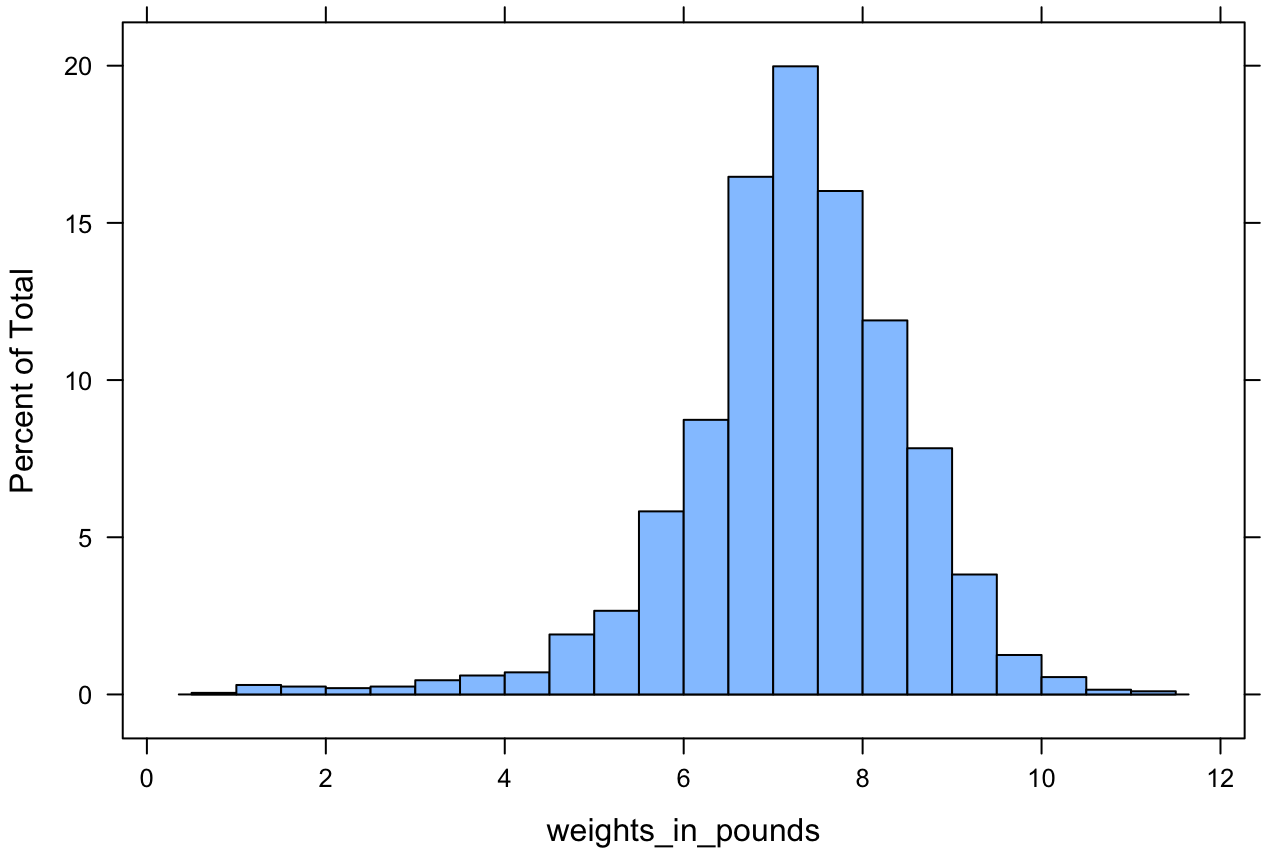
The CDC estimate is about 11.6% higher than the percentage from the NCbirths dataset.

6: Produce three different histograms of the weights in pounds. Use 3 bins, 20 bins, and 100 bins. Which histogram seems to give the best visualization, and why?

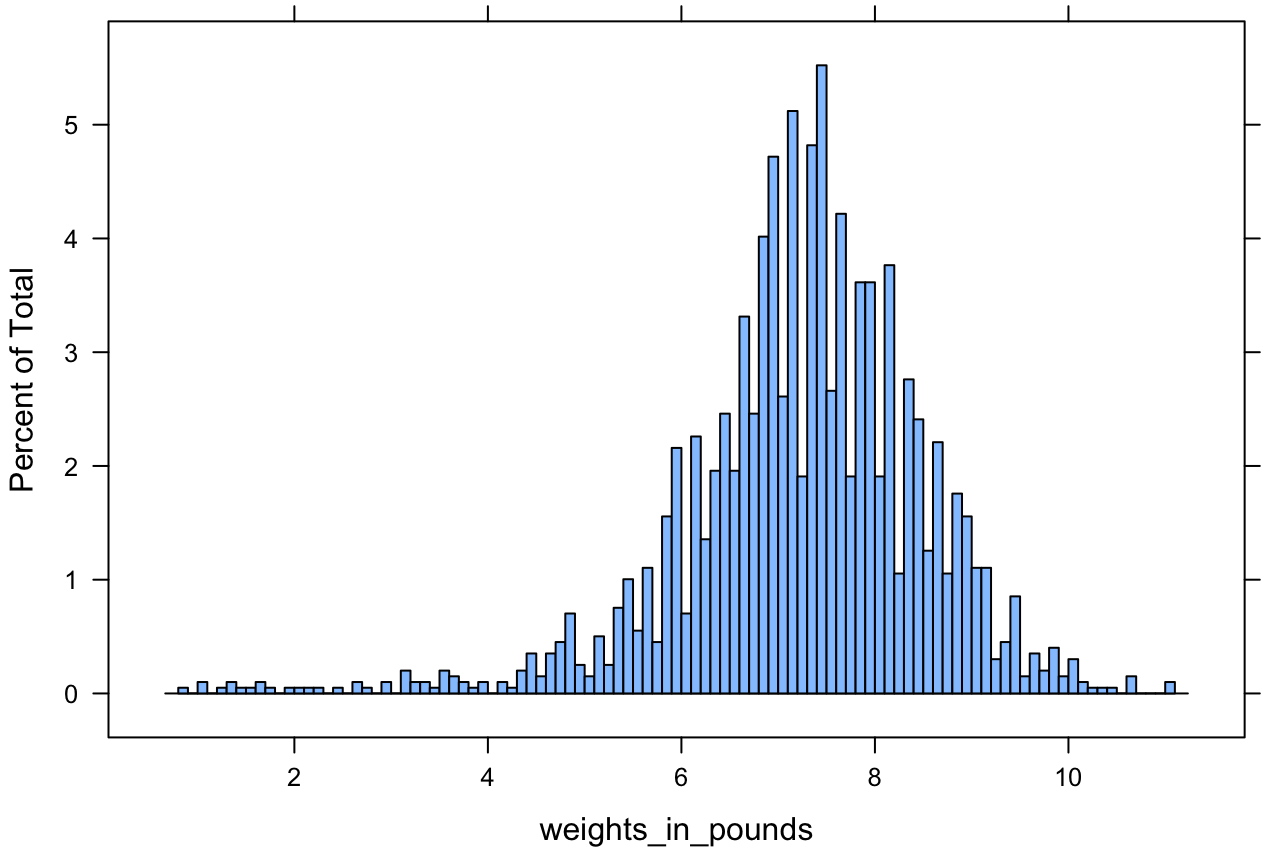
```
histogram(weights_in_pounds, breaks = 2)
```




```
histogram(weights_in_pounds, breaks = 19)
```




```
histogram(weights_in_pounds, breaks = 99)
```

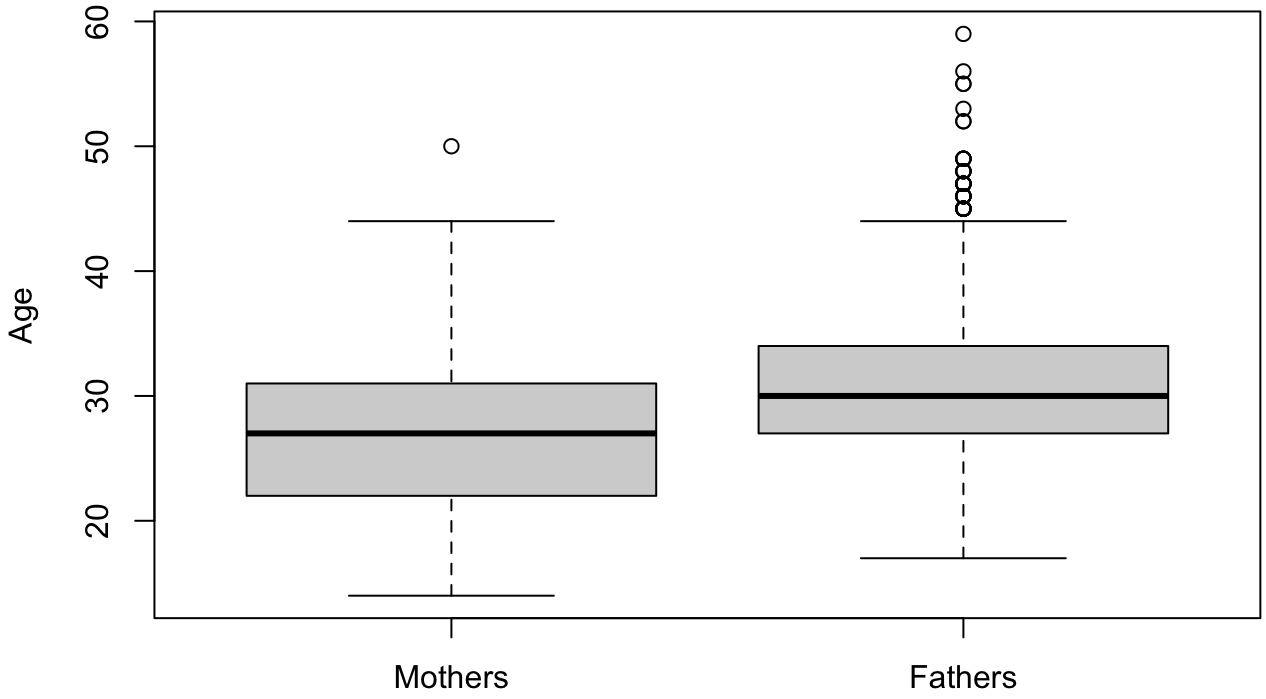


The histogram with 20 bins is a better visualization than the ones with 3 or 100 boxes. With 3 bins, too much of the data is grouped together in one bin. This means we can't see the overall shape of the data distribution and we can't tell if there are outliers or skewness. With 100 bins, the bins become so narrow that the histogram becomes more jagged. The overall trend of the data is harder to see when the bins become too narrow.

7: We can use the syntax boxplot(vector1, vector2) to make a side by side box plot. Create a side by side boxplot of the mother's ages and the father's ages. Which gender tends to be

older?

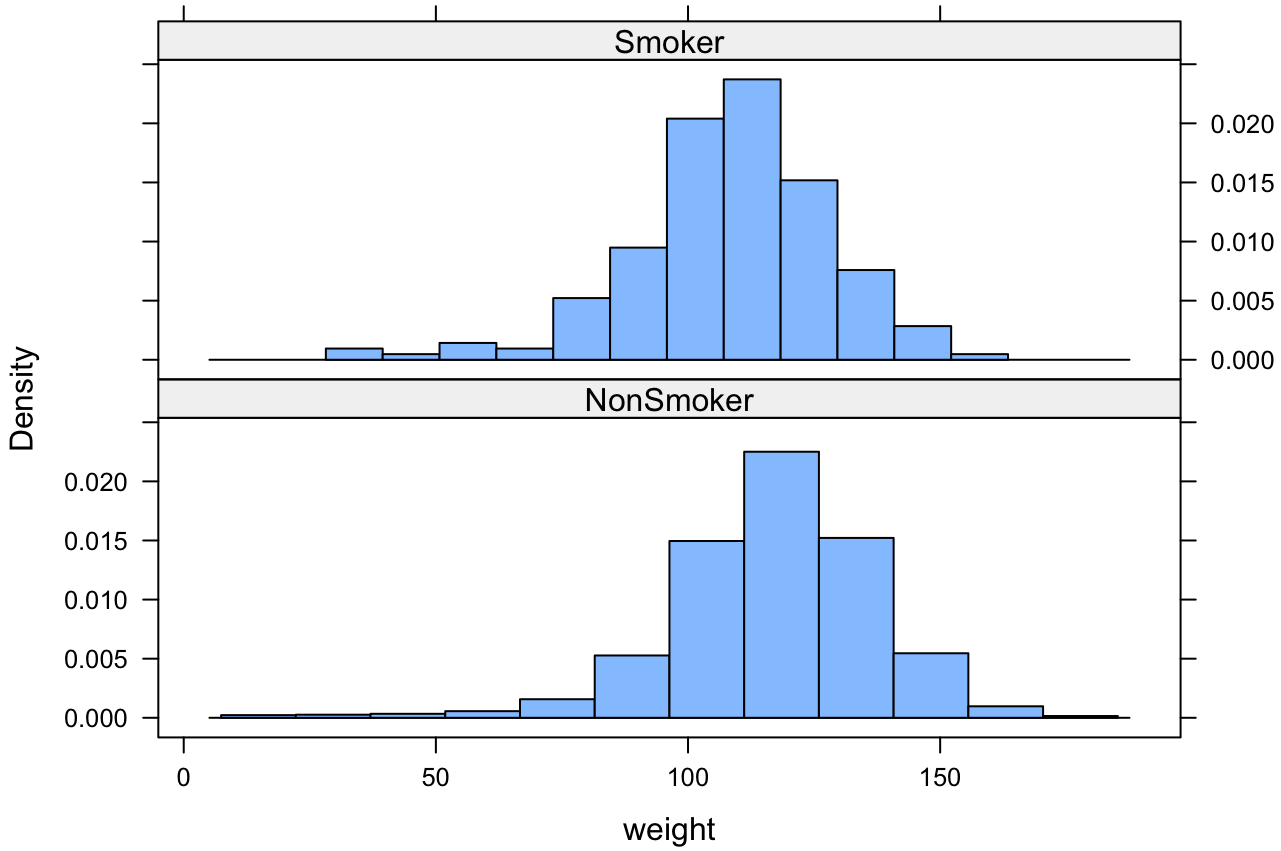
```
#Setting the y-label and names is optional
boxplot(NCbirths$Mage, NCbirths$Fage, ylab = "Age", names = c("Mothers","Fathers"))
```



Based on the boxplots, the fathers tend to be older than the mothers. Despite having roughly the same upper whisker, the boxplot for fathers has higher lower whisker, Q1, median, and Q3 values. Also, there are more potential outliers past the upper whisker on the fathers' boxplot compared to the mothers' plot.

8: Try typing histogram(~ weight | Habit, data = NCbirths, layout = c(1, 2)). Describe what this code does. Based on the graph, do you see any major differences between baby weights from smoking moms vs. non-smoking moms?

```
histogram(~ weight | Habit, data = NCbirths, layout = c(1, 2))
```

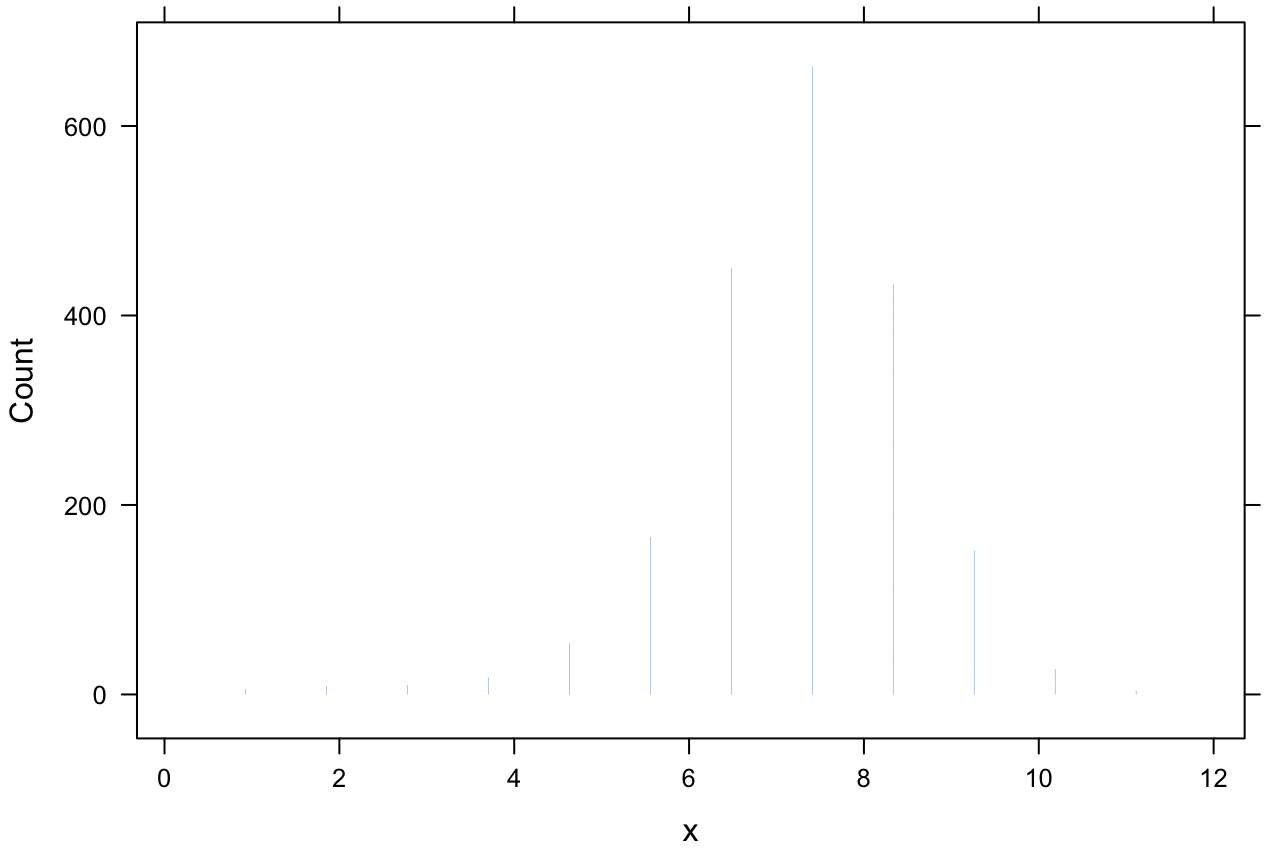


This code creates two histograms of the babies' weights, grouped by whether or not their mothers were smokers. The histogram for the smoker category seems to have a heavier lower tail, which could indicate a potential left-skew. Also, it seems like the center of the smoker histogram is slightly lower than the center of the non-smoker histogram. These details indicate that babies with non-smoking mothers are slightly heavier, on

average, than babies with non-smoking mothers.

9: Produce a dot plot of the weights in pounds.

```
dotPlot(weights_in_pounds)
```



10: Consider the other categorical variables in this data. Of those that record the health of the baby, which do you think will be associated with the mother's smoking and why? Make a two-way

Summary Table to check your hypothesis. Do you have evidence that this variable associated with smoking? Why?

The variables in the dataset that are related to the health of the babies are BirthDef (birth defects), DelivComp (delivery complications), and BirthComp (birth complications).I believe that babies with smoking mothers are more likely to have birth defects than babies with non-smoking mothers. The reason I believe this is because smoking is known to be carcinogenic. Additionally, other substances such as alcohol are known to cause birth defects such as Fetal Alcohol Syndrome. Thus, there might be a similar association between smoking and birth defects. To analyze this belief, I will examine BirthDef variable and its relationship with the Habit variable.

```
#You can use any format you want
two_way_table <- tally(~BirthDef | Habit, data = NCbirths, format = "proportion")
two_way_table
```

```
##              Habit
## BirthDef         NonSmoker      Smoker
##   At Least One 0.006648199 0.016042781
##   None         0.993351801 0.983957219
```

```
#Making a bar chart is optional
barplot(two_way_table, legend.text = TRUE)
```

As we can see in the two-way table and the bar chart, birth defects are quite rare across both the smoking and non-smoking groups. However, there are noticably more instances of birth defects in the smoking group. This indicates a possible relationship between smoking and birth defects. However, we need further testing to determine the strength of this relationship and the significance of the difference observed.

Note: The steps of this analysis will be the same for any of the three variables used. Regardless of the variable you picked, your conclusion and interpretation should be similar.

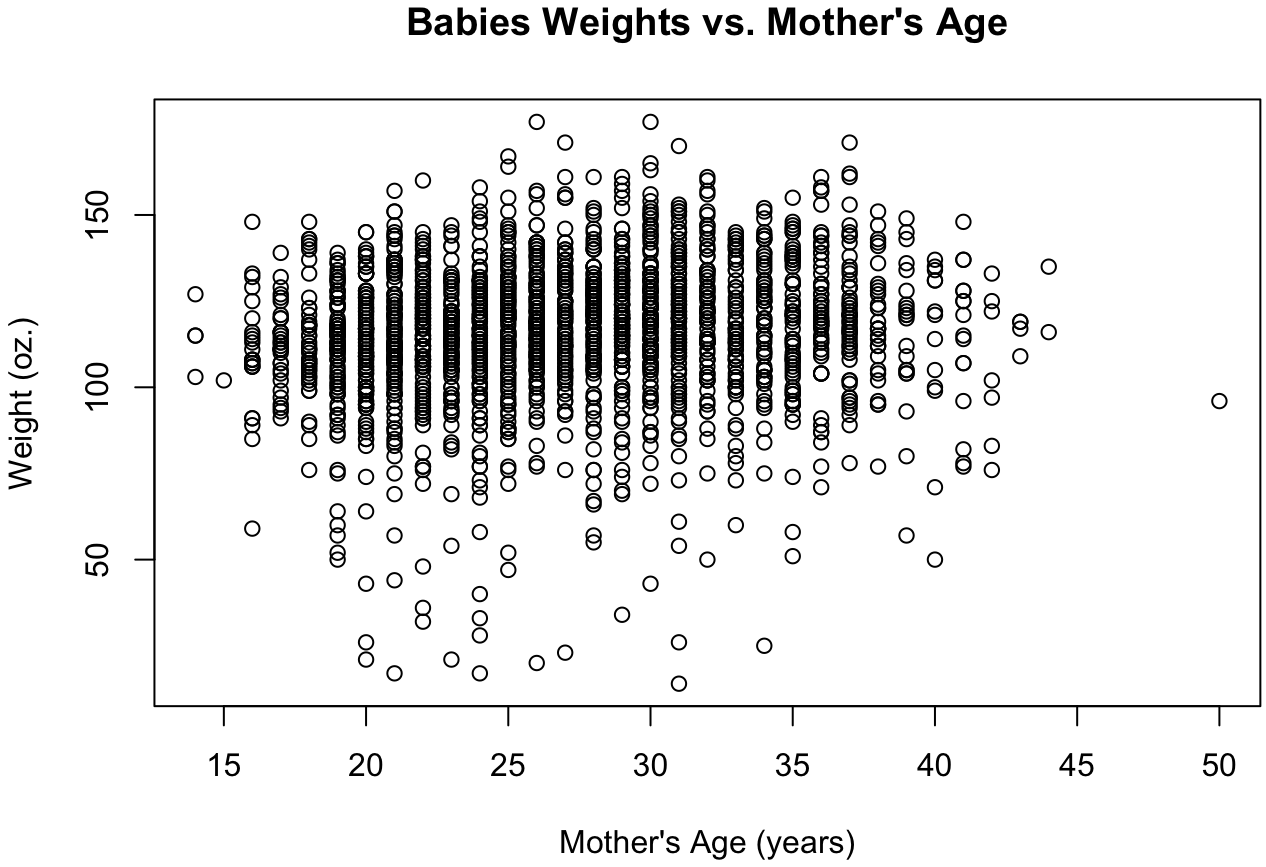
11: Produce a nicely formatted scatter plot of the weight of the

baby vs. the mother's age.

```
#plot(NCbirths$Mage, NCbirths$weight, xlab = "Mother's Age (years)", ylab = "Weight (oz.)",
#     main = "Babies Weights vs. Mother's Age")

#or

plot(weight ~ Mage, data = NCbirths, xlab = "Mother's Age (years)", ylab = "Weight (oz.)",
     main = "Babies Weights vs. Mother's Age")
```



Part 2

1: A data set on Shark Attacks Worldwide posted on StatCrunch records data on all shark attacks in recorded history including attacks before 1800. The data set can be viewed here: https://www.statcrunch.com/app/index.html?dataid=2188687 (https://www.statcrunch.com/app/index.html?dataid=2188687)

a) How many variables are contained in the data?

Each variable is represented as a column in the data set. Since there are 15 columns, there are 15 variables.

b) Which of the following questions could not be answered using this data set? Briefly explain.

i. In what month do most shark attacks occur?

ii. Are shark attacks more likely to occur in warm temperature or cooler temperatures?

iii. Attacks by which species of shark are more likely to result in a fatality?

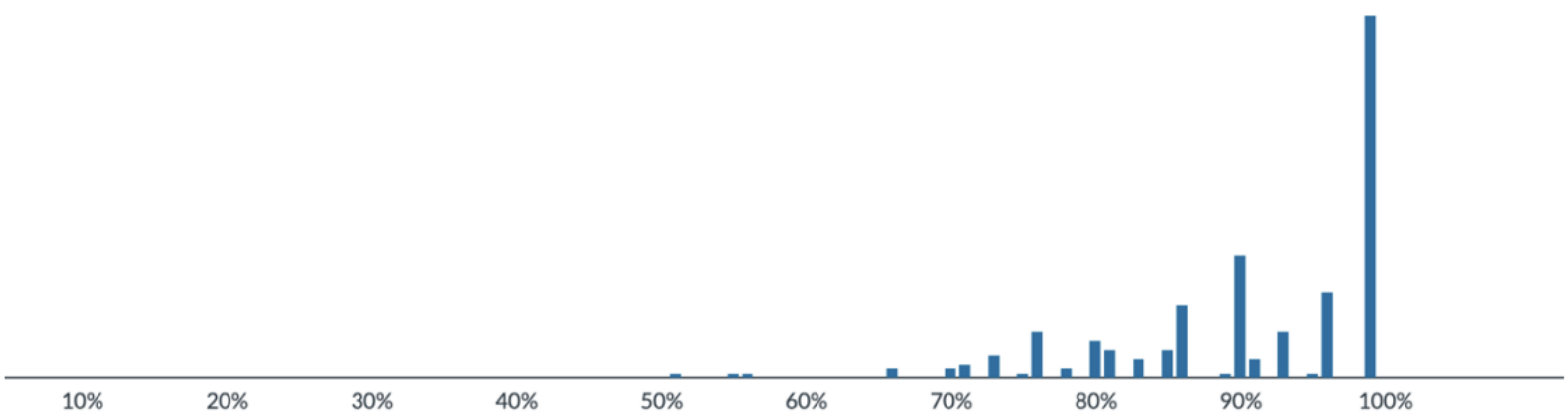
iv. What country has the most shark attacks per year?

The dataset has a variable for month, which allows us to answer question i. Similarly, there are variables that keep track of fatalities as well as the country an attack took place in. Thus, we can answer iii. and iv. In contrast, there are no temperature related variables, which means we cannot answer question ii.

c) A researcher wants to understand the age of the people in the data set and proposed some questions of interest: Are the reported cases are mostly younger people or older people? How is the age distributed? How would you help the research answer these questions? What statistical tools (e.g., graphs, measures) will you use? (You only need to describe your approach)

Since age is a numerical variable, I would first create a histogram of the ages in the dataset. From there, I can determine whether the data is skewed towards older or younger demographics. I can also identify potential outliers from the histogram. *Using other visualization tools such as a boxplot or dotplot are also acceptable answers*.

2: The scores of a quiz are displayed in the graph below.



a) Describe the shape of distribution

The histogram indicates that the data is unimodal and skewed to the left. There are potential outliers around 50%.

b) Would the mean score be greater than, less than, or about the same as the median score? Explain.

The mean score would be less than the median. Whenever the data is skewed, the mean is pulled in the direction of the skew. In this case, since the data is skewed to the left, the mean will be pulled to the left of the median.

c) What measures would you use to report the center and spread. Explain.

Since the data is skewed, the mean would not be an accurate measure of the center. Instead, we should use the median to measure the center of the data. Similarly, standard deviation is not a good measure of spread for skewed data. We should use IQR to measure the spread instead.

3: The distribution of test scores in a class is unimodal and symmetric with a mean of 80 pts and a standard deviation of 7pts. Based on the information, Adam estimated that his score is higher than approximately 97.5% of the students in class. What score did Adam receive? Explain.

Using the 68-95-99.7 rule, we know that 95% of the data lies between $\mu - 2\sigma$ and $\mu + 2\sigma$. The remaining 5% is split evenly between the right and left tail. Since Adam's score is higher than 97.5% of the data, his score is higher than the 2.5% in the left tail, as well as the 95% contained between $\mu \pm 2\sigma$. In other words, Adam's score is 2 standard deviations above the mean. The math below shows that Adam's score is 94.

$$\mu + 2\sigma = 80 + 2(7) = 94$$

Alternatively, we can use a Z-table to find the Adam's z-score. The z-score associated with an observation that's greater than 97.5% of the data is $z = 1.96$. From here, we can use the formula

$$x = \sigma z + \mu$$

to find Adam's score. Plugging in the quantities gives us

$$x = 7(1.96) + 80 = 93.72.$$

Either 94 or 93.72 are acceptable answers, as long as proper work is shown.

4: Assume that both men and women's heights have symmetric and unimodal distributions. Women's distribution has a mean of 64 inches and a standard deviation of 2.5 inches. Men's distribution has a mean of 69 inches and a standard deviation of 3 inches.

a) What women's height corresponds with a z-score of -1.50?

The formula for a z-score is given by

$$z = \frac{x - \mu}{\sigma},$$

where x is the observed data point, μ is the mean, and σ is the standard deviation. Solving for x gives us

$$x = \sigma z + \mu.$$

Since we are looking at the women's data, we will use $\mu = 64$ and $\sigma = 2.5$. Plugging these in, along with $z = -1.5$, gives us

$$\begin{aligned} x &= \sigma z + \mu \\ &= 2.5(-1.5) + 64 \\ &= 2.5(-1.5) + 64 \\ &= 60.25. \end{aligned}$$

Thus, a z-score of -1.5 corresponds to a women's height of 60.25 inches.

b) Professional basketball player Evelyn Akhator is 75 inches tall and plays in the WNBA (women's league). Professional basketball player Draymond Green is 79 inches tall and plays in the NBA (men's league). Compared to their own peers, who is taller?

To compare Evelyn Akhator's and Draymond Green's heights, we need to standardize them by computing their respective z-scores. Starting with Evelyn, we will find the z-score using the women's mean and standard deviation:

$$\begin{aligned} z_{\text{Evelyn}} &= \frac{x - \mu}{\sigma} \\ &= \frac{75 - 64}{2.5} \\ &= 4.4. \end{aligned}$$

Now, for Draymond, we will use the mean and standard deviation of the men:

$$\begin{aligned} z_{\text{Draymond}} &= \frac{x - \mu}{\sigma} \\ &= \frac{79 - 69}{3} \\ &\approx 3.3. \end{aligned}$$

Notice that the z-score for Draymond Green is less than the z-score for Evelyn Akhator. This means that Evelyn is taller compared to her peers than Draymond is compared to his peers.

5: The top ten movies based on Marvel comic book characters for the U.S. box office as of fall 2017 are shown in the following table, with domestic gross rounded to the nearest hundred

million. (Source: ultimatemovieranking.com)

Movie	Domestic Gross ($ millions)
The Avengers (2012)	677
Spiderman (2002)	602
Spiderman 2 (2004)	520
Avengers: Age of Ultron (2015)	471
Iron Man 3 (2013)	434
Spiderman 3 (2007)	423
Captain America: Civil War (2016)	408
Guardians of the Galaxy Vol. 2 (2017)	389
Iron Man (2008)	384
Deadpool (2016)	363

a) Report the five-number summary of the domestic gross income.

In my answer, all numbers will be in terms of millions of dollars. Based on the table, we can see that the minimum value is 363 and the maximum value is 677. The data is already organized from greatest to least, so the median will be the middle entry. Since there are 10 entries, the median will be the average between the 5th and 6th entries.

$$\text{Median} = \frac{423 + 434}{2} = 428.5$$

To find the Q1 and Q3 values, we can split the data into two subsets. One subset will include all of the data greater than the median, and the other will include all of the entries below the media. From there, we can find the median of each of these subsets. The median of the upper subset is 520 and the medain of the lower half is 389. Therefore, our five number summary is:

- Min = 363
- Q1 = 389
- Median = 428.5
- Q3 = 520
- Max = 677

b) Interpret the five-number summary in context, i.e., what information can you obtain about the distribution of the domestic gross income?

Based on the five number summary, we can measure the center and spread of the data. The median, which is a measure of the center, is \$428.5M. So, a typical Marvel movie in the top ten will gross roughly \$428.5M. The IQR, which is \$520M - \$389M = \$131M, can be used to measure the spread of the data. Specifically, we know that the center 50% of the data will be spread out over \$131M. Finally, we can compare the distances between the numbers in the five number summary to check for skew and potential outliers. It seems like the distance between the Max and Q3 is much larger than the distance between the Min and Q1. Also, it appears as though the median is closer to Q1 than Q3. This indicates that the data might have a right-skew.

6: The data set below show the number of central public libraries

in 32 states.

States	Number of Central Libraries	States	Number of Central Libraries
Connecticut	182	Colorado	113
Vermont	155	New Hampshire	219
Oregon	129	Washington	62
Hawaii	1	Mississippi	52
Idaho	102	South Dakota	112
Montana	82	Louisiana	68
New Jersey	281	Nevada	21
Georgia	63	Alaska	79
Alabama	218	New York	756
Texas	548	Kentucky	119
Indiana	237	Virginia	91
District of Columbia	1	Arkansas	58
Utah	72	Massachusetts	368
Ohio	251	Rhode Island	48
South Carolina	42	Florida	82
North Dakota	73		

The five number summary is given as:

- Minimum = 1
- Q1 = 62

- Median = 91
- Q3 = 218
- Maximum = 756

Sketch a boxplot using the five-number summary above and the data below. Mark the values of the quartiles, the lower whisker, the upper whisker, and any potential outliers in the boxplot. Explain how you determined the length of the whiskers.

Before sketching the boxplot, we need to calculate the upper and lower whiskers. The upper and lower bounds for the whiskers can found using the formulas:

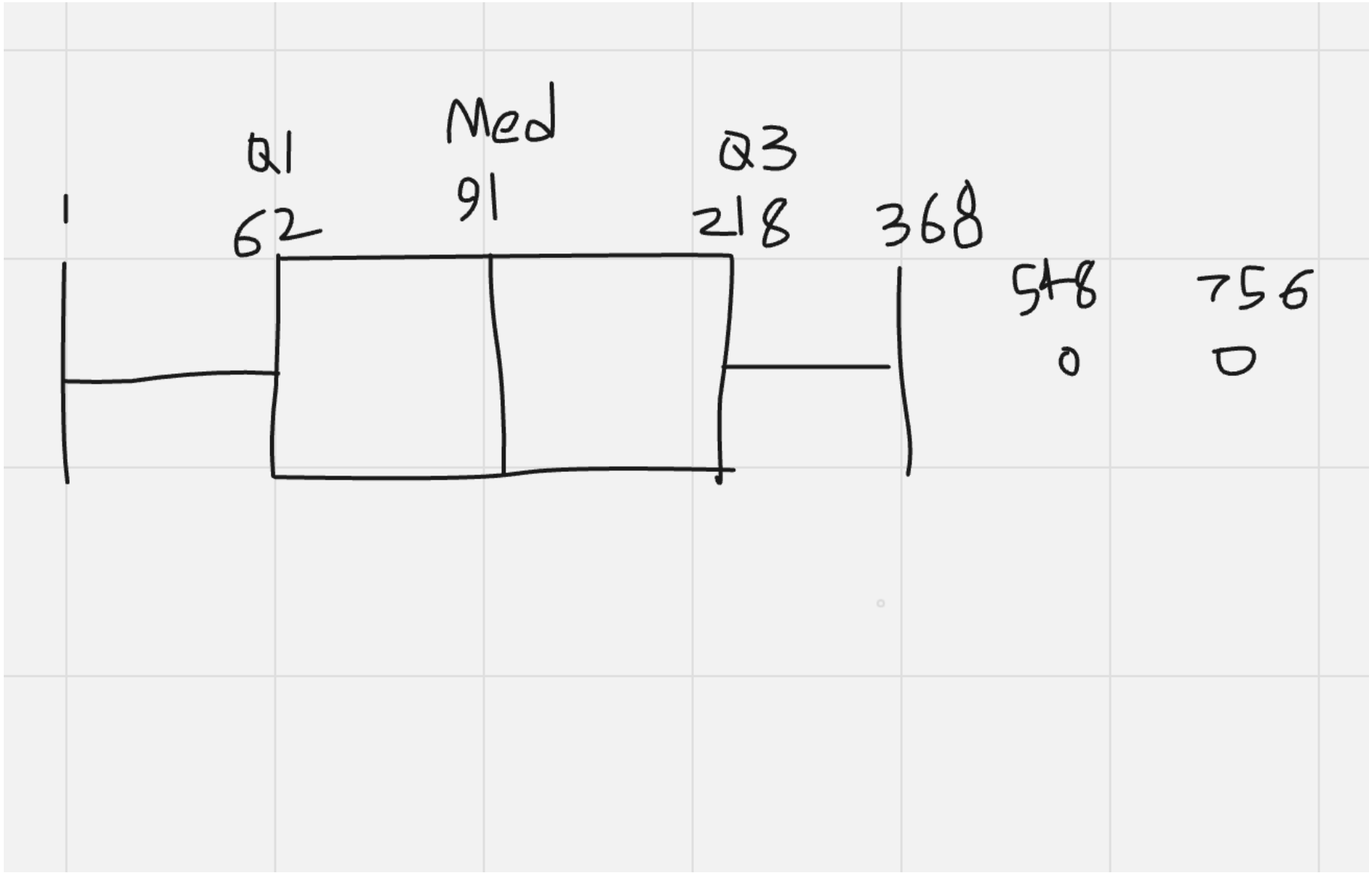
- $\text{LowerBound} = \text{Q1} - 1.5 \cdot \text{IQR}$
- $\text{UpperBound} = \text{Q3} + 1.5 \cdot \text{IQR}$

The IQR is given by $218 - 62 = 156$. Plugging Q1,Q3, and the IQR into the bounds formulas gives us

$$\text{LowerBound} = 62 - 1.5(156) = -172$$
$$\text{UpperBound} = 218 + 1.5(156) = 452.$$

To find the actual upper whisker value, we have to find the largest data point that is lower than the upper bound. Looking at the table, it is clear that 368 is the largest value that does not exceed the upper bound. Similarly, the lower whisker value is the smallest data point that is above the lower bound. Clearly, the minimum value of 1 is the smallest data point that is above the lower bound.

Note that we have two data points that are above the upper bound. These are New York at 756 and Texas at 548. These two points are potential outliers that we will represent as dots in our boxplot.



Potential Boxplot