Homework 1

Welcome to the first homework assignment!

The purpose of this homework is to practice using R and R Markdown, and to review some concepts from introductory Statistics. Please fill in the appropriate R and R Markdown and write answers to all questions in the answer section., then submit a compiled pdf or html with your answers to Canvas by 11:59pm on Sunday September 8th.

If you need help with any of the homework assignments, please attend the TA office hours which are listed on Canvas and/or ask questions on Piazza. Also, if you have completed the homework, please help others out by answering questions on Piazza.

Problem 1: RMarkdown practice

RMarkdown has a number of features that allow the text in your written reports to have better formatting. In the following exercise, please modify lines of text to change their formatting. A cheatsheet for RMarkdown formatting can be found here. When answering the questions (i.e., formatting the text below) be sure to knit your RMarkdown document very often to catch errors as soon as they are made.

Problem 1.1: Please format the lines of text below (15 points)

Make this line bold

Make this line italics

Make this line a third level header

- Make this line a bullet point

LINK

Problem 1.2: Use LaTeX to write plato's name in Greek below (5 points)

Note: make sure the ending dollar sign touches the last letter otherwise you will get an error when knitting.

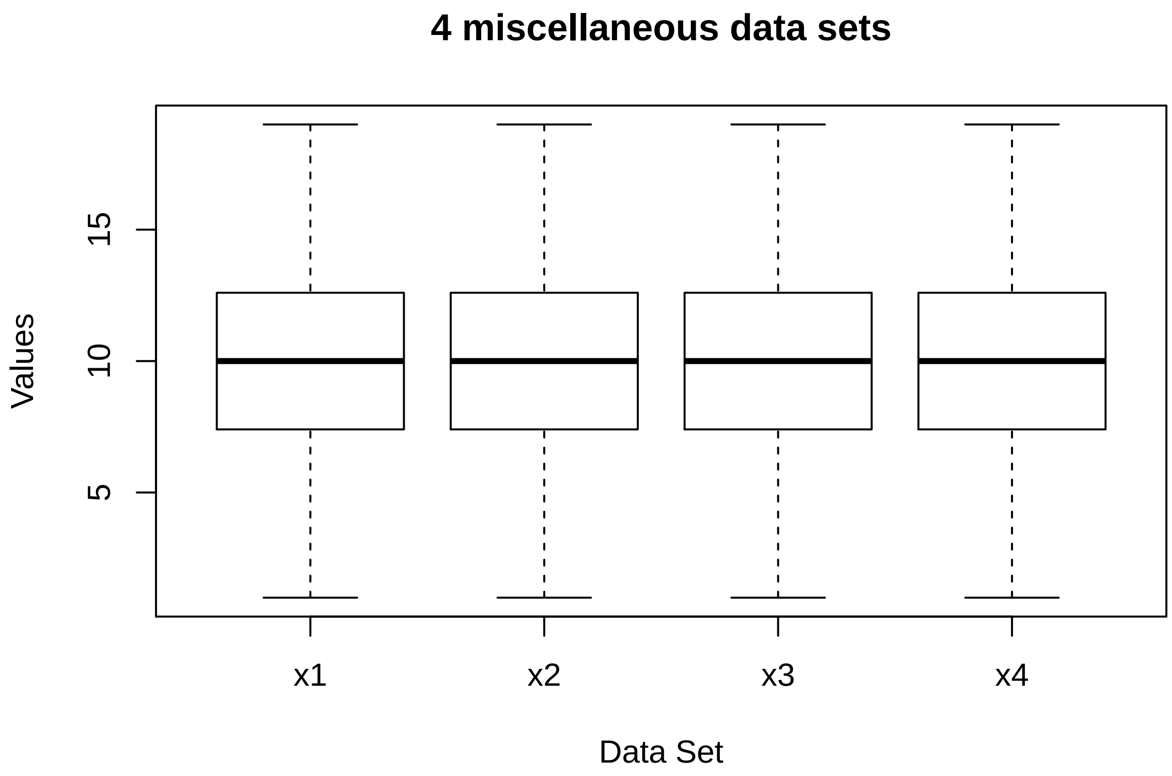
$\Pi\lambda\acute{\alpha}\tau\omega\upsilon$

Problem 2: Descriptive statistics and plots

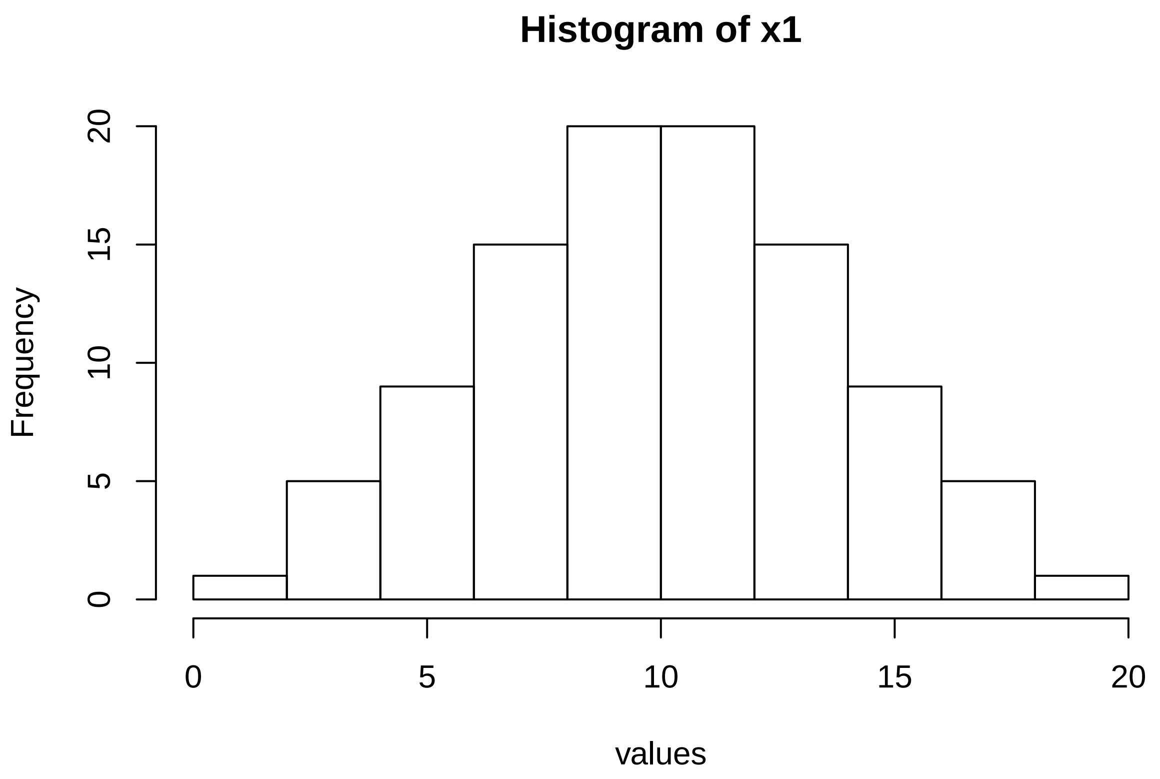
Below you will create and compare a few plots. Please answer each question, and if you notice any outliers in your data please address them appropriately. Also be sure to label your plots appropriately.

Part 2.1: (10 points) The code chunk below loads four vector objects named x1, x2, x3, and x4. Create a side-by-side boxplot that compares these four vectors. Also create a histogram for each of these vectors (4 histograms total). Describe below whether the boxplots or histograms are more informative for plotting this data and why.

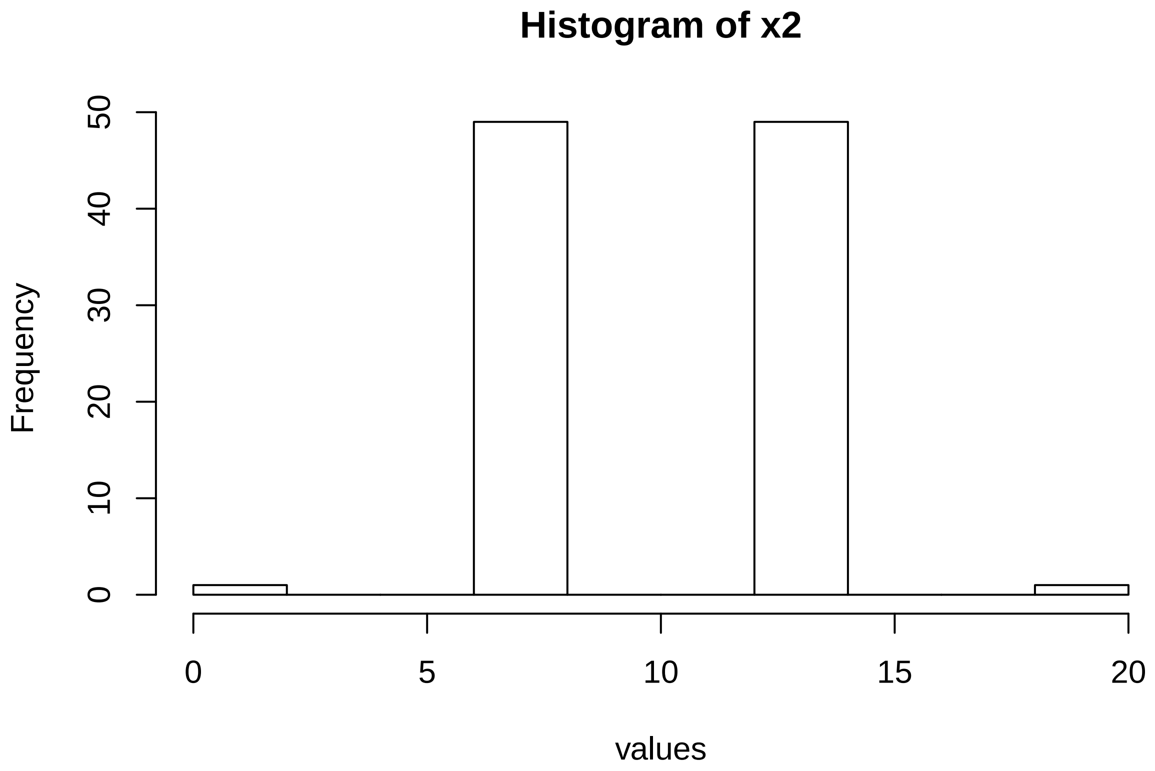
```r
load("misc_data.Rda")

boxplot(x1, x2, x3, x4, xlab = "Data Set",
        ylab = "Values", main = "4 miscellaneous data sets",
        names = c("x1", "x2", "x3", "x4"))
```



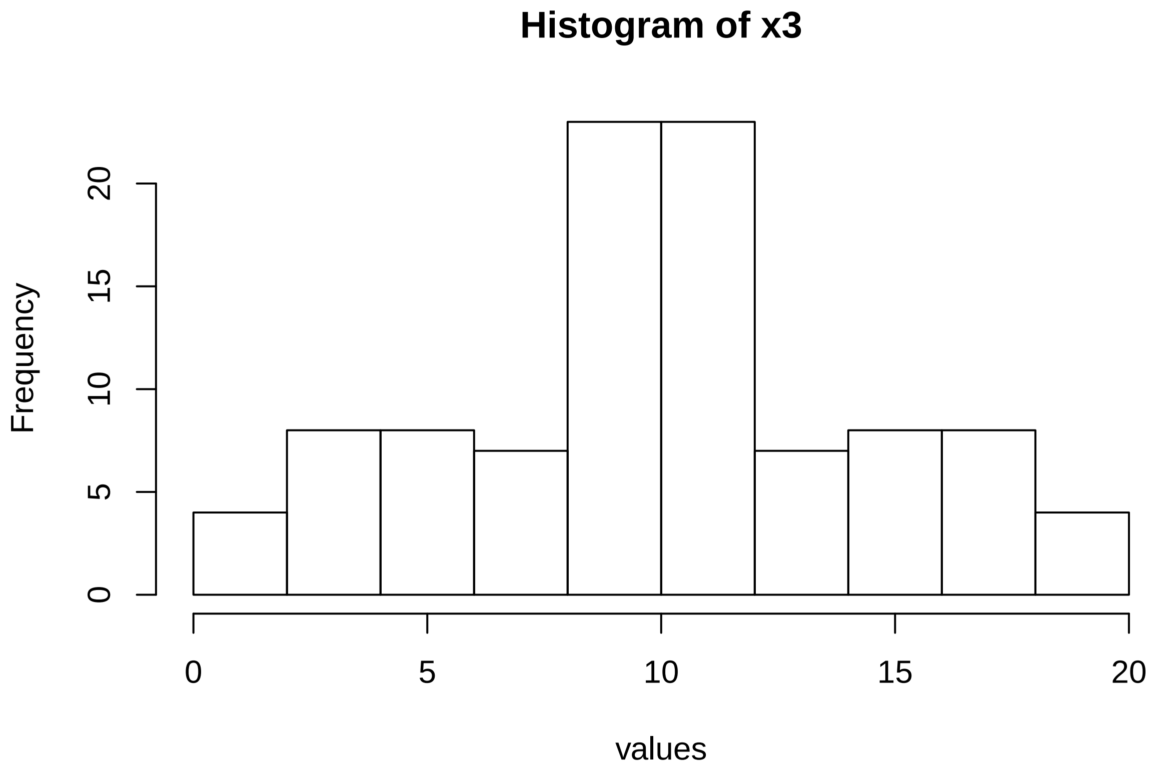
```r
hist(x1, xlab = "values")
```



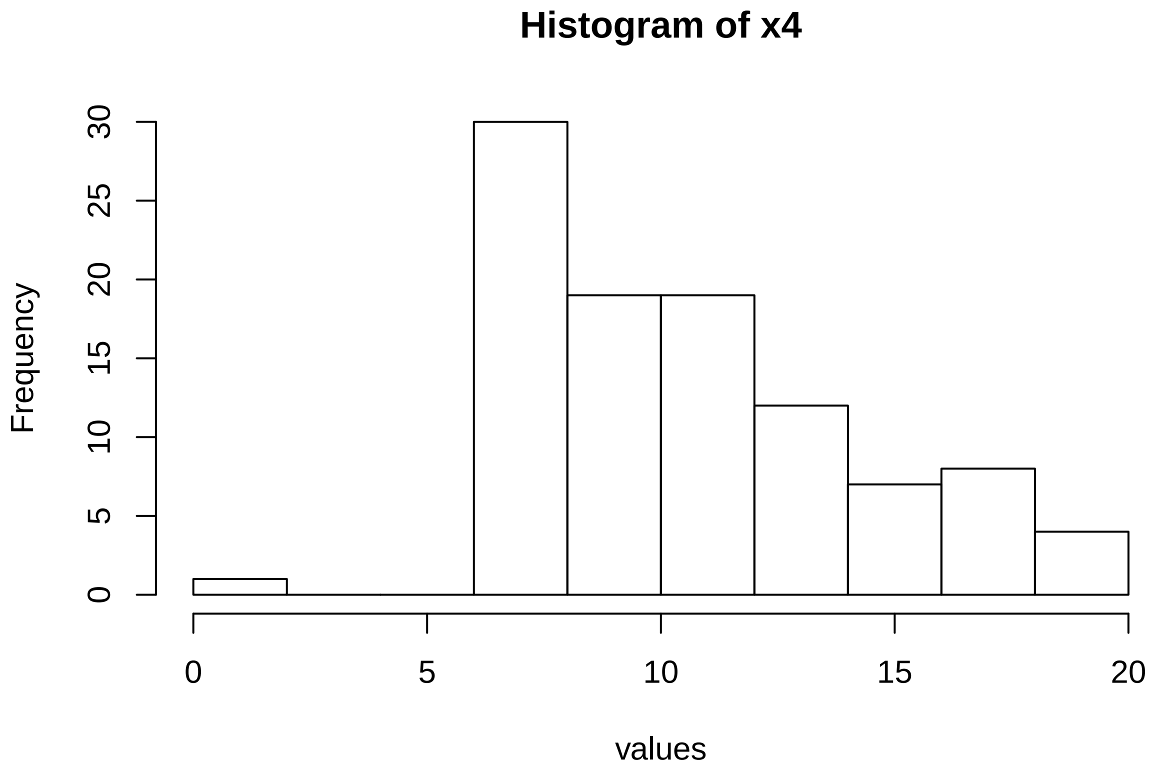
```r
hist(x2, xlab = "values")
```



```r
hist(x3, xlab = "values")
```



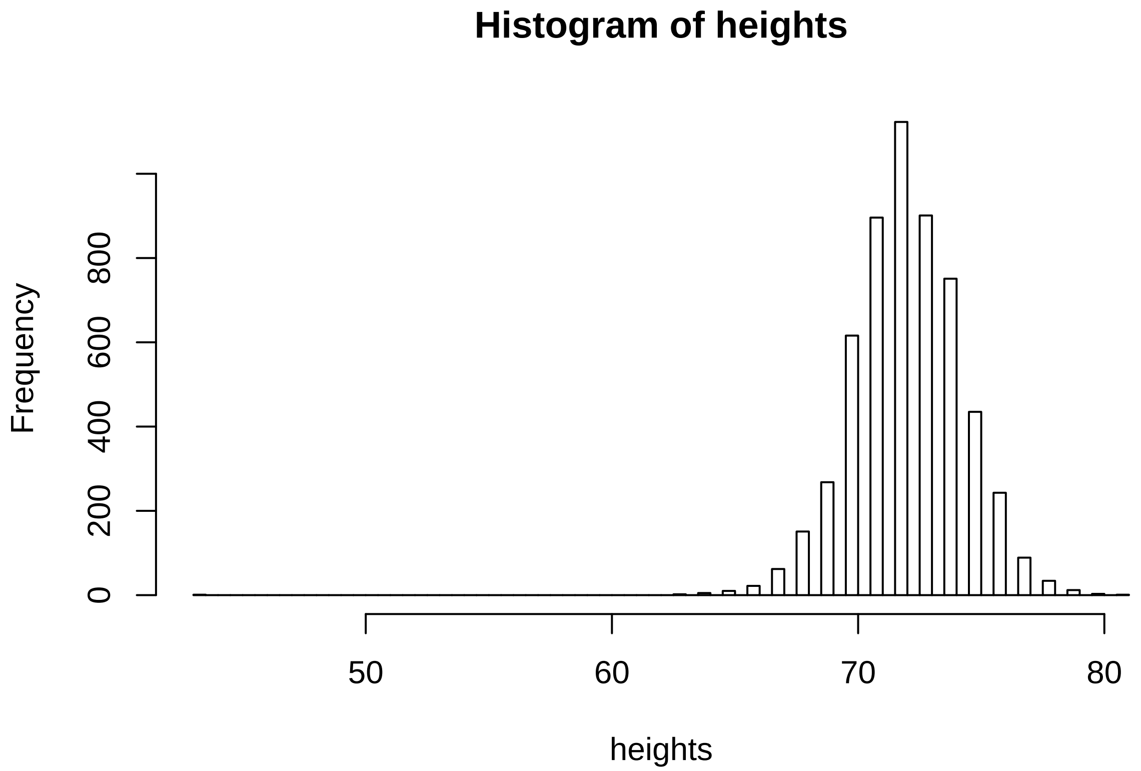
```r
hist(x4, xlab = "values")
```



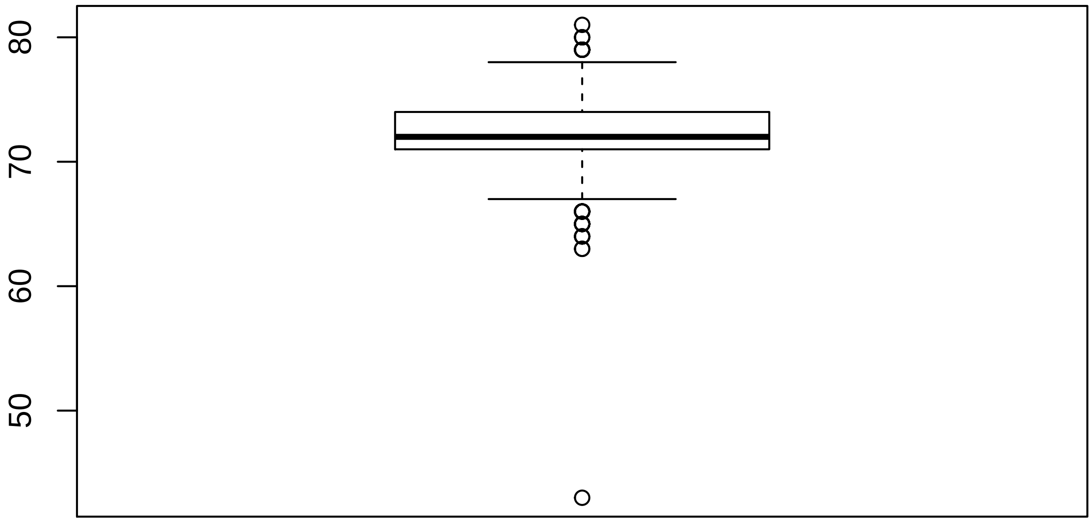
Answer: [Describe whether boxplots or histograms are more informative here] Clearly the histograms are much more informative than boxplots since all the boxplots are identical, even though the data is quiet different. This illustrates an important limitation of boxplots.

Part 2.2: (10 points) The R chunk below loads a data frame with information on all major league baseball players who were born between 1901 and 1950 (if you are interested in the data, it comes from the Lahman package). Create a vector object that is called `heights`, that has just the player heights. Then create a histogram and a boxplot of the players' heights using this vector object. Describe the shape of the distribution of heights and any advantages that one type of plot has over the other. Also investigate any unusual features of the data.

```r
load("players_born_1901_1950.Rda")
heights <- players_born_1901_1950$height

hist(heights, nclass = 100)
```



```
boxplot(heights)
```

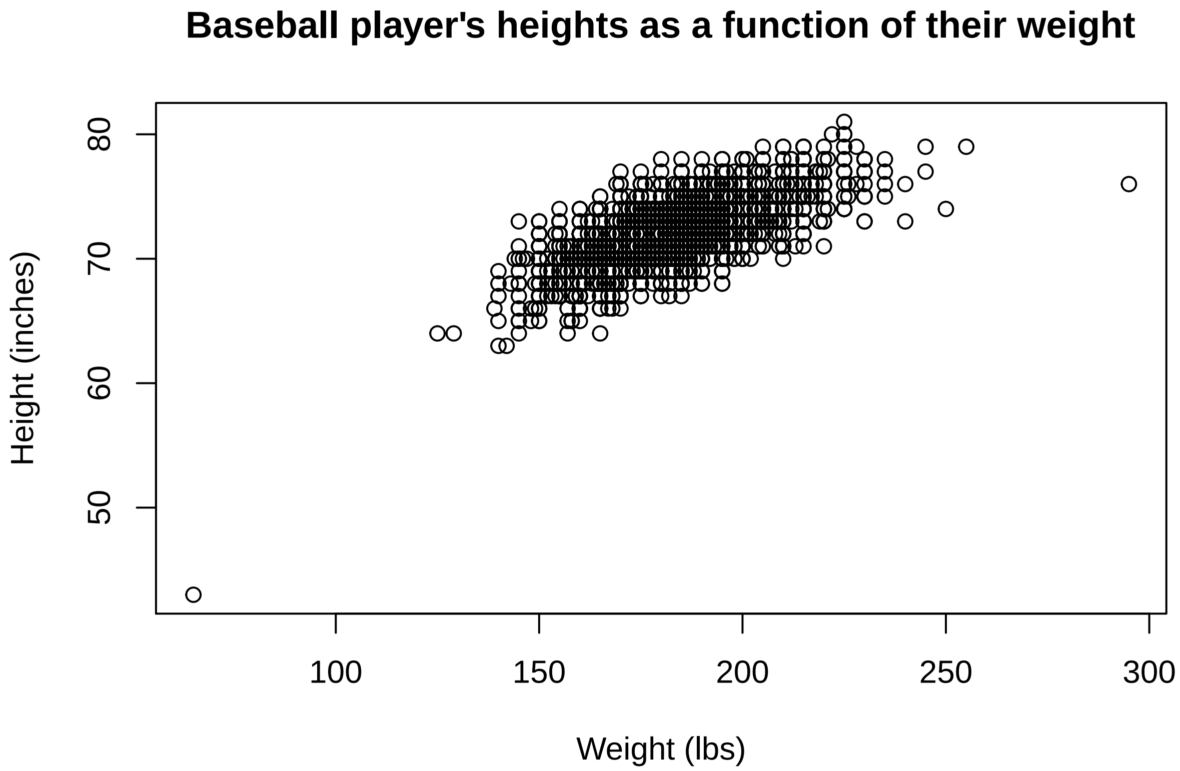


Answer: [Describe advantages of boxplots and histograms for this data and investigate usual features of the data]. The histogram shows that the distribution of the heights seems to be normal. The boxplot reveals that there is one extreme outlier. The advantage of the histogram is that we can see the shape of the full distribution while the advantage of the boxplot is that it makes it clear there is an extreme outlier.

Investigating this outlier by googling ‘short baseball player’ reveals that this player is Eddie Gaedel. More information about Eddie can be found on his wikipedia entry.

Part 2.3: (10 points) Create a scatter plot of the baseball player’s heights as a function of their weight. Describe what the results show.

```r
plot(players_born_1901_1950$weight, players_born_1901_1950$height,
     xlab = "Weight (lbs)", ylab = "Height (inches)",
     main = "Baseball player's heights as a function of their weight")
```



Answer: The scatter plot shows a positive association between players weights and their heights, i.e., taller players tend to weigh more. This is what I would expect.

Problem 3: Examining categorical data

Let's now examine which states/regions baseball players are born in.

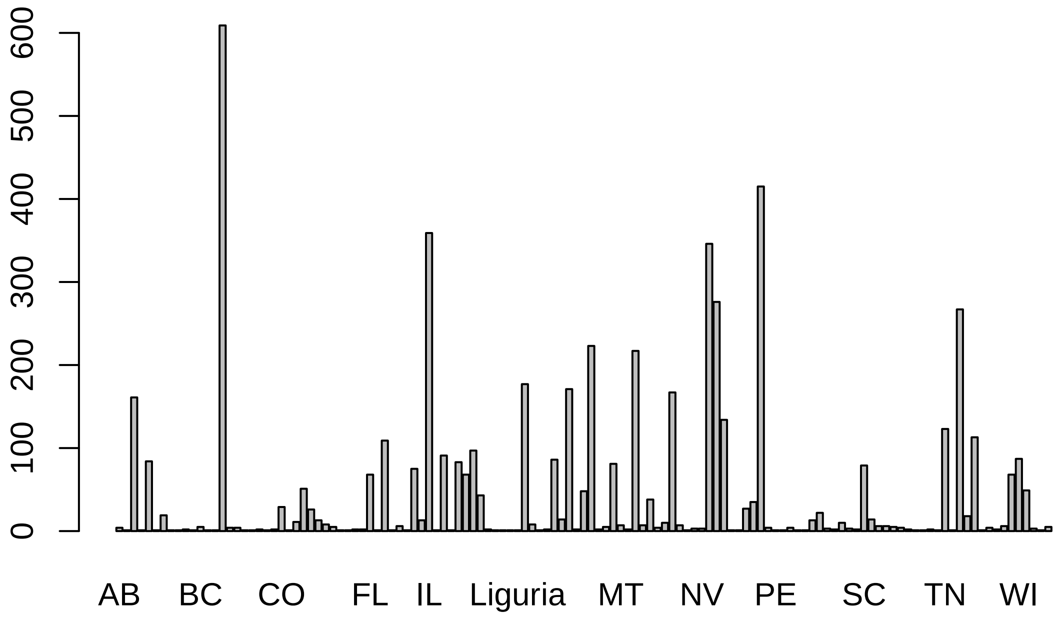
Part 3.1: (10 points) Use the table() function to create an object called birth_place_counts that has the counts of where players were born in. What is the state that the most players were born in?

Then create a bar plot and pie chart showing the counts of places that players are born in. How do these plots look? How could we make them better?

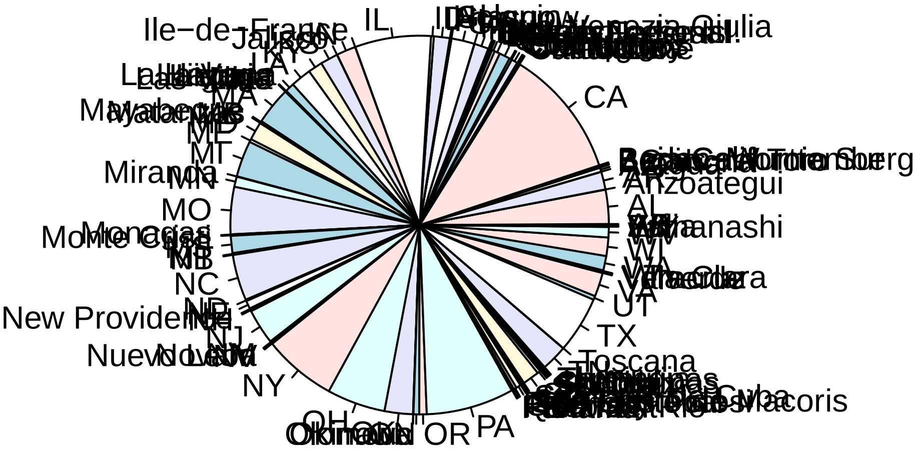
```
birth_place_counts <- table(players_born_1901_1950$birthState)

sort(birth_place_counts)
```

```
## 
##                    AK            Anzoategui                Aragua 
##                     1                     1                     1 
##     Baden-Wurttemberg       Baja California              Barahona 
##                     1                     1                     1 
##                Berlin        Bocas del Toro              Carabobo
##                     1                     1                     1
##              Cheshire              Chiriqui              Coahuila
##                     1                     1                     1
##      Dodescanese Isl.                Duarte Friuli-Venezia Giulia
##                     1                     1                     1
##               Glasgow               Holguin         Ile-de-France
##                     1                     1                     1
##               Jalisco                  Lara            Las Villas
##                     1                     1                     1
##               Liepaja               Liguria             Mayabeque
##                     1                     1                     1
##                Novara               Okinawa               Olomouc
##                     1                     1                     1
##                    PE              Piedmont              Plzensky
##                     1                     1                     1
##                Puebla                 Sucre               Suffolk
##                     1                     1                     1
##             Thuringia               Toscana              Valverde
##                     1                     1                     1
##             Yamanashi   Baja California Sur             Chihuahua
##                     1                     2                     2
##            Cienfuegos              El Seibo                Falcon
##                     2                     2                     2
##               La Vega                    MB               Miranda
##                     2                     2                     2
##               Monagas                    NB       San Luis Potosi
##                     2                     2                     2
##      Santiago de Cuba            St. Thomas            Tamaulipas
##                     2                     2                     2
##           Villa Clara            Nuevo Leon                    NV
##                     2                     3                     3
##         San Cristobal              Santiago                    WY
##                     3                     3                     3
##                    AB              Camaguey            Canal Zone
##                     4                     4                     4
##        New Providence                Panama         Pinar del Rio
##                     4                     4                     4
##             St. Croix              Veracruz                    BC
##                     4                     4                     5
##     Distrito Nacional          Monte Cristi                Sonora
##                     5                     5                     5
##                 Zulia                    HI               Sinaloa
##                     5                     6                     6
##                    SK                    VT                    MT
##                     6                     6                     7
##                    ND                    NM      Distrito Federal
##                     7                     7                     8
##              Matanzas                    NH  San Pedro de Macoris
##                     8                    10                    10
##                 Colon                    DE                    ID
##                    11                    13                    13
##                    QC                    ME                    SD
##                    13                    14                    14
##                    UT                    AZ                    RI
##                    18                    19                    22
##                    DC                    ON                    CO
##                    26                    27                    29
##                    OR                    NE             La Habana
##                    35                    38                    43
##                    MN                    WV                    CT
##                    48                    49                    51
##                    FL                    KY                    WA
##                    68                    68                    68
##                    IA                    SC                    MS
##                    75                    79                    81
##                    KS                    AR                    MD
##                    83                    84                    86
##                    WI                    IN                    LA
##                    87                    91                    97
##                    GA                    VA                    TN
##                   109                   113                   123
##                    OK                    AL                    NJ
##                   134                   161                   167
##                    MI                    MA                    NC
##                   171                   177                   217
##                    MO                    TX                    OH
##                   223                   267                   276
##                    NY                    IL                    PA
##                   346                   359                   415
##                    CA
##                   609
```

```r
barplot(birth_place_counts)
```



```
pie(birth_place_counts)
```

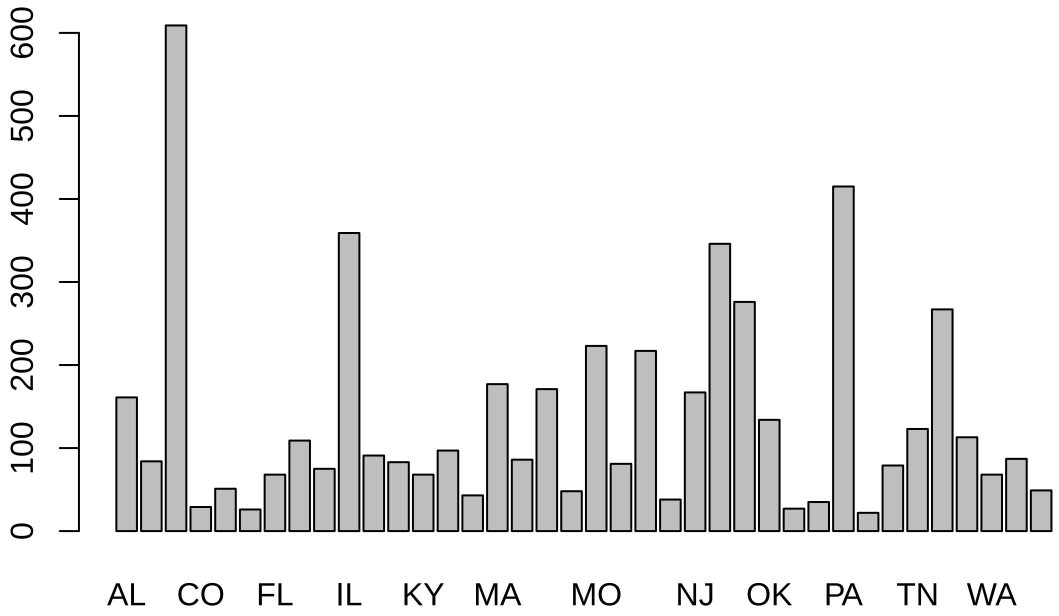


Answers: The place where most players were born in is California. These plots are ugly because they have too many places (particularly where few people were born). The colors are also pretty ugly.

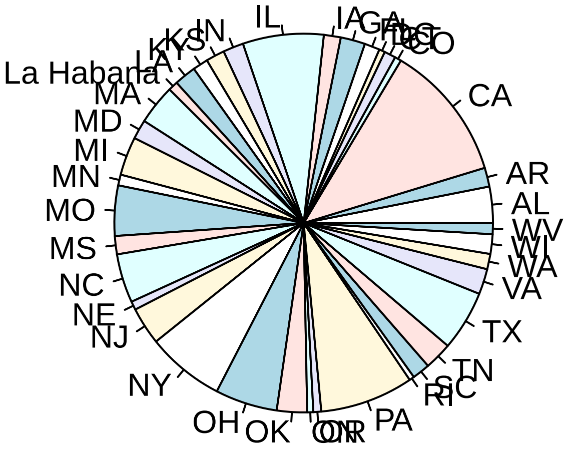
Part 3.2: (10 points)

Let's only plot states/places that have more than 20 players born in them. You can do this by creating a vector of booleans where TRUE indicates a state that has greater than 20 players born in it and FALSE indicates that 20 or less players were born in it (this can be done in 1 line of code). Then use this vector to extract only the places which more than 20 players born in. Finally replot the results with only states with more than 20 players born in them. Does this look better? Is there any place on this list that is not a state?

```
inds_counts_greater_20 <- birth_place_counts > 20

birth_place_counts_greater_20 <- birth_place_counts[inds_counts_greater_20]

barplot(birth_place_counts_greater_20)
```



```
pie(birth_place_counts_greater_20)
```

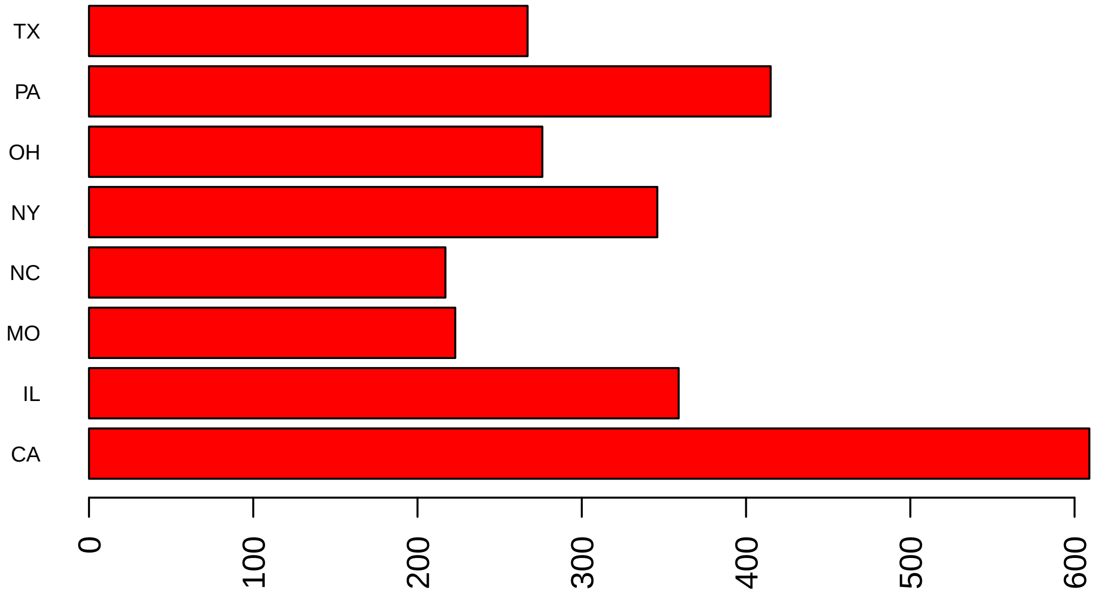


Answer: Yes better, but still there is still too much stuff. The place on the initial list that was not a state is Habana.

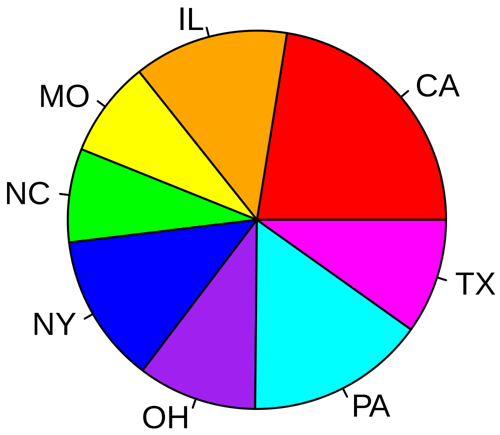
Part 3.3: (10 points)

The plots in part 3.2 still could look better. Adjust the plots so that you plot fewer states so that it is easier to see exactly which states most players are born in. Also adjust other visual attributes of the plots so that none of the labels are overlapping, and see if you can find other ways to make the plots look better, e.g., by adjusting the colors, etc. (hint: using ? pie and google will be helpful). Is plotting only some of the states misleading in any way, and if so, what are ways this could be addressed?

```
inds_counts_greater_200 <- birth_place_counts > 200

birth_place_counts_greater_200 <- birth_place_counts[inds_counts_greater_200]

barplot(birth_place_counts_greater_200, horiz = TRUE, cex.names = .7, col = "red", las=2)
```



```
colors = c("red", "orange", "yellow", "green", "blue", "purple", "cyan", "magenta")

pie(birth_place_counts_greater_200, col = colors)
```



Answer: Only plotting some of the states can be misleading, particularly for the pie chart, because it gives the sense that a proportion of the players come from a particular state when this proportion is only out of the states that have over 200 players born in them (e.g., it looks like ~25% of players come from CA when this is not the case). A way to address this would be to create a cateogry called "other state" that has the number of players born in all states that are not shown.

Problem 4: For loops (10 points)

As discussed in class, for loops allow you to repeat a process many times. Each time the process is repeated, a counter index object (usually named *i*) is incremented by 1. This is useful because it allows you to:

1. Repeat a process many times to generate results each time
2. Store each result in a vector using *i* to index into the vector.

The code below create uses a for loop to store the values of 1 squared up to 50 squared in a vector object named my_vec. Modify the code so that what is stored in the vector are the even integers from 2 to 100 (i.e, 2, 4, 6, . . . , 100).

```
my_vec <- NULL  # need to initialize the vector so R knows what to put the results into
for (i in 1:50){
  my_vec[i] <- 2 * i
}

my_vec
```

```
##  [1]   2   4   6   8  10  12  14  16  18  20  22  24  26  28  30  32  34
## [18]  36  38  40  42  44  46  48  50  52  54  56  58  60  62  64  66  68
## [35]  70  72  74  76  78  80  82  84  86  88  90  92  94  96  98 100
```

Problem 5: Short reading (5 points)

As discussed in class, OkCupid is a dating website. One of the founders of the website, Christian Rudder, created a series of blog posts around 2010 where he analyzed data from the site to extract insights about dating. In order gain insight into what is possible from simple descriptive statistics and plots, please read the blog entry from July 7th 2010 title 'The Big Lies People Tell In Online Dating' and write one paragraph comment on something interesting you found in the article. Alternatively, you can read and comment on the article title 'How a Math Genius Hacked OkCupid to Find True Love' and comment on that article instead.

Describe something interesting you found in one of these articles:

Reflection (5 points)

Please reflect on how the homework went. In particular, please answer the following questions:

1. What concepts do you feel you are clearly understanding and which concepts are you confused about?
2. How many hours did you spend working on the homework?
3. How much did you enjoy doing the homework ("Super fun", "kind of fun", "not really", or "terrible")?
4. How much do you feel you learned doing this homework ("learned a lot", "learned some", "learned nothing", or "even more confused")?
5. Please note also if you went to TA office hours for help with this worksheet, and if the help you got was useful (in general, we strongly encourage you to attend TA office hours if you are having any difficulties with the homework).
6. Anything else you would like us to know?

Reflection Answers:

1.

2.

3.

4.

5.

6.