STATS 10 Assignment 1

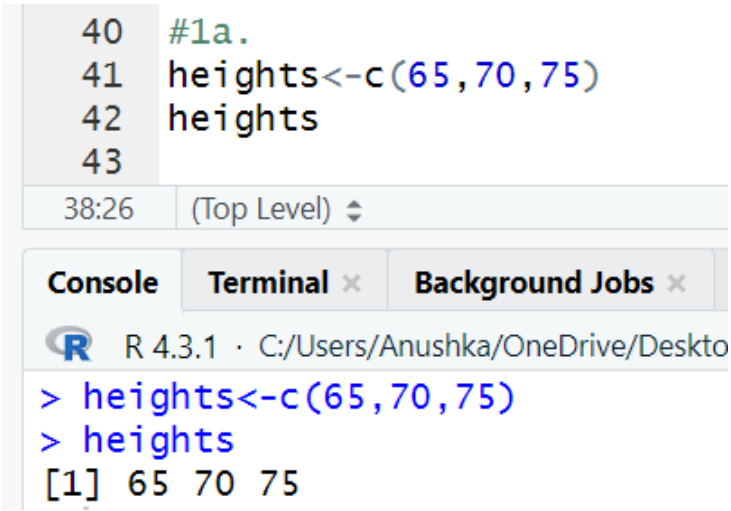
Anushka Nayak

UID: 605977416

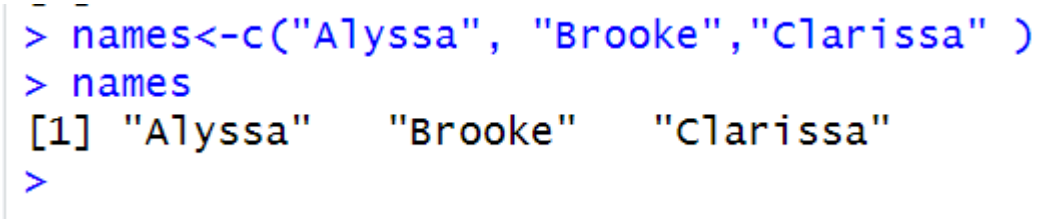
PART I

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.



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



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

```
> cbind(heights, names)
     heights names
[1,] "65"    "Alyssa"
[2,] "70"    "Brooke"
[3,] "75"    "Clarissa"
> class(cbind(heights, names))
[1] "matrix" "array"
```

This command combines the vectors into a matrix. Class of the new object is matrix.

2. Downloading data:

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

```
> NCbirths<- read.csv("births.csv")
> class(NCbirths)
[1] "data.frame"
```

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   Marit
al
1   Male     No    124      8   31   25    13    14       1     13   Marri
ed
2 Female     No    177      8   36   26     9    12       2     11 Unmarri
ed
3   Male     No    107      3   30   16    12     8       2     10 Unmarri
ed
4 Female     No    144      6   33   37    12    14       2     12 Unmarri
ed
5   Male     No    117      9   36   33    10    16       2     19   Marri
ed
6 Female     No     98      4   31   29    14    16       3     20   Marri
ed
  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. Package loading

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

```
> library(maps)
> find.package("maps")
[1] "C:/Users/Anushka/AppData/Local/R/win-library/4.3/maps"
>
```

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

Use the births data set for questions 4-11

4. Perform vector operations

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

```
> weight <- NCbirths$weight
> weight
   [1] 124 177 107 144 117  98 147 138 104 123 153 129 119
  [14] 108 106 125 115 128 132  83 117 130 130 103  85 133
  [27] 122 134  84 117 118 164 147 106 144 117  95 112 115
  [40] 107 105 119 143 112 177 119 136 119  33 118 134 106
  [53] 118 106 130 112 102 134 116 134 117  61 132 119 129
  [66]  57 130 104 118 123 135 124 118  77 128  94 122 108
  [79] 116 117 115 112  89 113 122  83 111  72 151 125 114
```

 [92] 141 109 93 96 117 116 86 156 137 86 125 150 142
[105] 112 136 105 155 98 119 106 109 114 139 112 121 97
[118] 85 121 138 76 140 115 91 120 142 116 132 122 99
[131] 128 114 97 132 127 77 122 113 132 115 74 122 117
[144] 136 111 118 143 110 122 109 120 92 121 156 141 109
[157] 130 112 87 112 125 112 100 108 110 111 119 118 93
[170] 107 132 130 138 142 95 100 107 129 114 108 131 95
[183] 83 136 128 116 103 151 110 134 145 112 118 96 117
[196] 154 102 108 130 131 117 85 95 93 122 108 120 91
[209] 119 135 117 95 145 34 136 84 111 110 131 138 115
[222] 103 60 113 104 107 36 136 109 120 107 135 148 122
[235] 119 111 104 139 116 131 120 138 113 148 145 133 117
[248] 142 99 105 143 118 105 123 120 125 116 102 120 94
[261] 131 136 124 117 128 126 120 111 126 71 104 133 126
[274] 131 121 106 111 131 96 128 103 135 131 101 128 135
[287] 104 116 116 103 137 125 127 84 136 108 136 109 119
[300] 138 132 106 140 113 112 113 113 113 124 118 134 131
[313] 118 114 98 165 125 114 122 119 117 120 121 113 88
[326] 109 126 121 128 121 103 120 123 119 109 113 113 130
[339] 111 138 109 145 112 121 122 147 99 113 114 89 103
[352] 120 123 122 113 91 108 77 109 114 113 144 87 128
[365] 109 98 104 115 112 50 69 115 120 134 105 128 117
[378] 100 80 116 131 48 85 126 108 124 54 111 131 112
[391] 123 114 124 134 80 85 119 140 106 120 92 143 111
[404] 135 105 145 127 122 109 73 109 97 98 57 96 102
[417] 107 124 133 118 126 113 80 112 104 115 89 145 131
[430] 99 123 101 133 76 120 132 90 122 118 54 108 121
[443] 123 110 122 129 135 118 144 91 128 102 112 137 129
[456] 116 122 133 133 122 111 137 87 50 145 112 125 120
[469] 120 106 128 138 131 96 113 113 47 107 127 134 108
[482] 131 96 108 133 127 132 125 87 99 99 104 137 108
[495] 76 112 125 103 115 126 153 125 120 147 111 102 125
[508] 112 130 149 104 70 147 120 108 160 123 148 112 112
[521] 72 102 115 101 136 110 141 121 93 137 108 100 127
[534] 139 120 105 108 137 134 131 103 111 84 141 161 111
[547] 133 105 158 136 107 132 109 109 119 123 115 125 97
[560] 151 116 112 95 122 121 124 121 114 122 123 135 123
[573] 111 142 92 124 124 139 106 109 117 104 124 144 113
[586] 119 136 137 89 117 83 132 99 145 122 115 83 127
[599] 100 109 124 121 111 108 106 112 98 153 123 138 129
[612] 139 128 170 139 151 116 140 116 119 157 111 96 110
[625] 115 117 112 100 113 120 115 119 74 156 138 106 121
[638] 114 113 144 129 124 87 145 151 137 43 135 115 133
[651] 145 126 119 143 140 109 96 120 104 105 93 134 92
[664] 111 117 147 134 112 110 87 128 127 142 128 147 130
[677] 161 83 129 123 129 138 146 130 112 109 117 108 102
[690] 100 102 95 107 127 112 127 96 81 102 108 125 131
[703] 120 120 130 142 124 95 136 108 137 120 108 117 99
[716] 171 155 95 51 40 146 114 139 106 110 96 135 125
[729] 122 107 115 104 119 120 107 93 118 99 115 100 157
[742] 101 138 112 126 110 114 98 143 126 102 136 125 104
[755] 69 131 121 96 138 89 103 121 110 77 111 102 138
[768] 98 122 119 137 117 121 124 114 111 125 138 140 102
[781] 133 109 113 105 99 113 149 127 120 120 129 87 106
[794] 114 78 106 111 126 147 85 108 123 118 133 114 94
[807] 135 115 110 111 131 126 146 117 117 114 150 138 129
[820] 121 108 116 113 99 104 123 114 127 131 111 113 143
[833] 136 115 157 108 112 119 126 108 152 91 122 126 103
[846] 119 119 137 132 106 113 117 129 106 75 124 123 135
[859] 77 116 137 119 85 115 105 17 124 126 121 88 116
[872] 130 139 128 104 112 104 116 93 142 76 118 118 101
[885] 118 118 122 123 126 98 114 114 129 114 152 120 120
[898] 105 43 103 121 116 99 119 117 131 118 137 104 86
[911] 108 109 114 89 105 137 141 151 130 117 119 113 114
[924] 76 140 102 119 133 105 107 137 124 32 136 115 102
[937] 130 94 105 139 109 114 117 71 91 121 103 116 133
[950] 120 100 133 151 128 132 110 96 100 141 121 142 112
[963] 113 115 93 120 117 72 130 93 122 115 91 96 141
[976] 131 117 131 77 107 143 110 152 141 100 138 123 114

```
 [989] 101 121  98 136  82 151 128 118 141 122  55 100
 [ reached getOption("max.print") -- omitted 992 entries ]
```

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

The weight is measured 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 = weight*0.0625
> weights_in_pounds
   [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  7.3125  8.1250  8.1250  6.4375  5.3125  8.3125  7.6250
  [28]  8.3750  5.2500  7.3125  7.3750 10.2500  9.1875  6.6250  9.0000  7.3125
  [37]  5.9375  7.0000  7.1875  6.6875  6.5625  7.4375  8.9375  7.0000 11.0625
  [46]  7.4375  8.5000  7.4375  2.0625  7.3750  8.3750  6.6250  7.3750  6.6250
  [55]  8.1250  7.0000  6.3750  8.3750  7.2500  8.3750  7.3125  3.8125  8.2500
  [64]  7.4375  8.0625  3.5625  8.1250  6.5000  7.3750  7.6875  8.4375  7.7500
  [73]  7.3750  4.8125  8.0000  5.8750  7.6250  6.7500  7.2500  7.3125  7.1875
  [82]  7.0000  5.5625  7.0625  7.6250  5.1875  6.9375  4.5000  9.4375  7.8125
  [91]  7.1250  8.8125  6.8125  5.8125  6.0000  7.3125  7.2500  5.3750  9.7500
 [100]  8.5625  5.3750  7.8125  9.3750  8.8750  7.0000  8.5000  6.5625  9.6875
```

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
```

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.

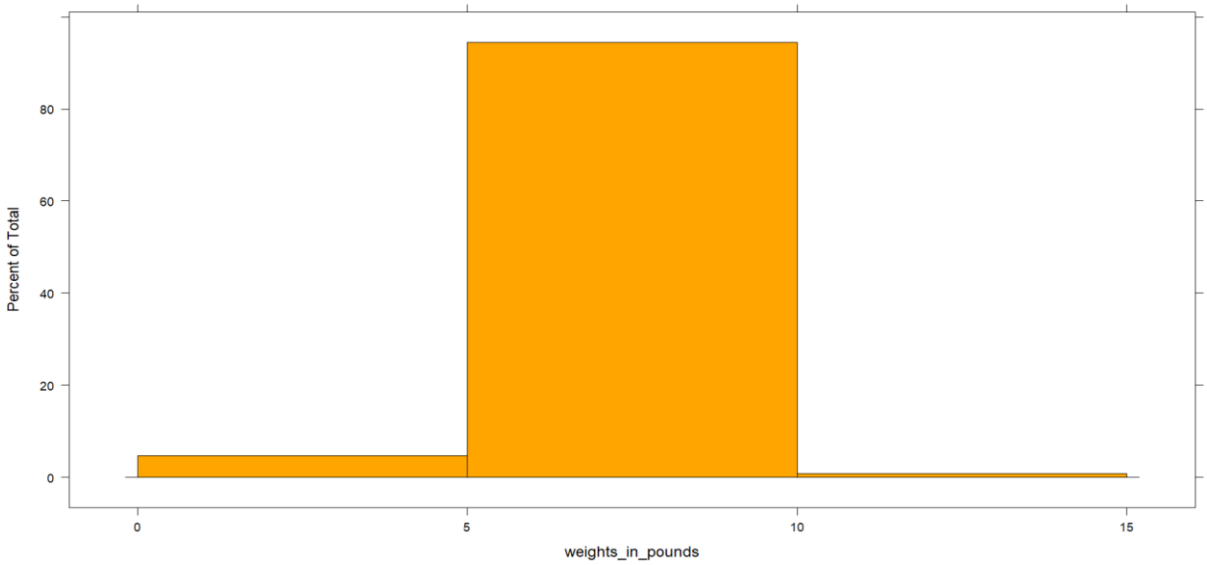
```
> library(mosaic)
>  tally(NCbirths$Habit, format="percent")
X
NonSmoker     Smoker
 90.61245    9.38755
```

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?

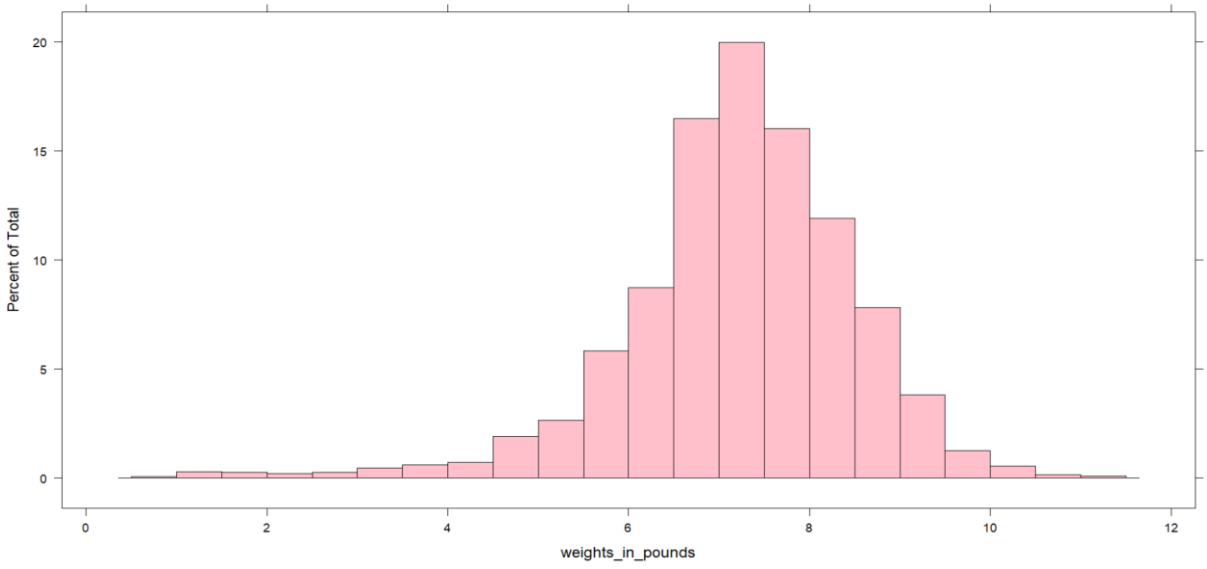
The percentage in question 2 is 11.612450% far off from the CDC's report.

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?

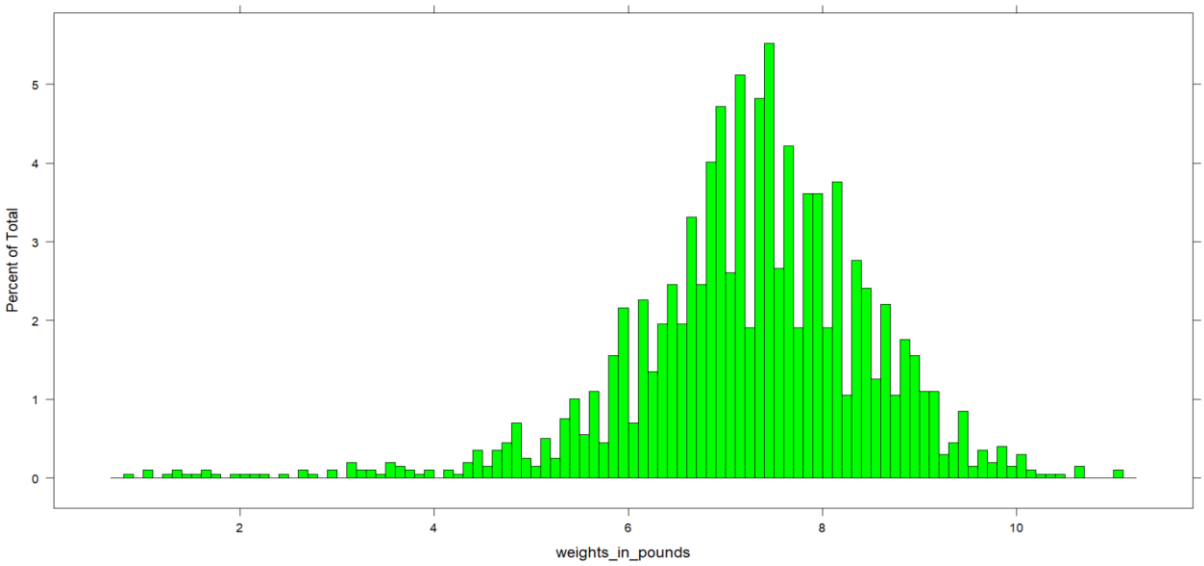
```
> histogram(weights_in_pounds, breaks=3, col="orange")
```




```
> histogram(weights_in_pounds, breaks=20, col="pink")
```



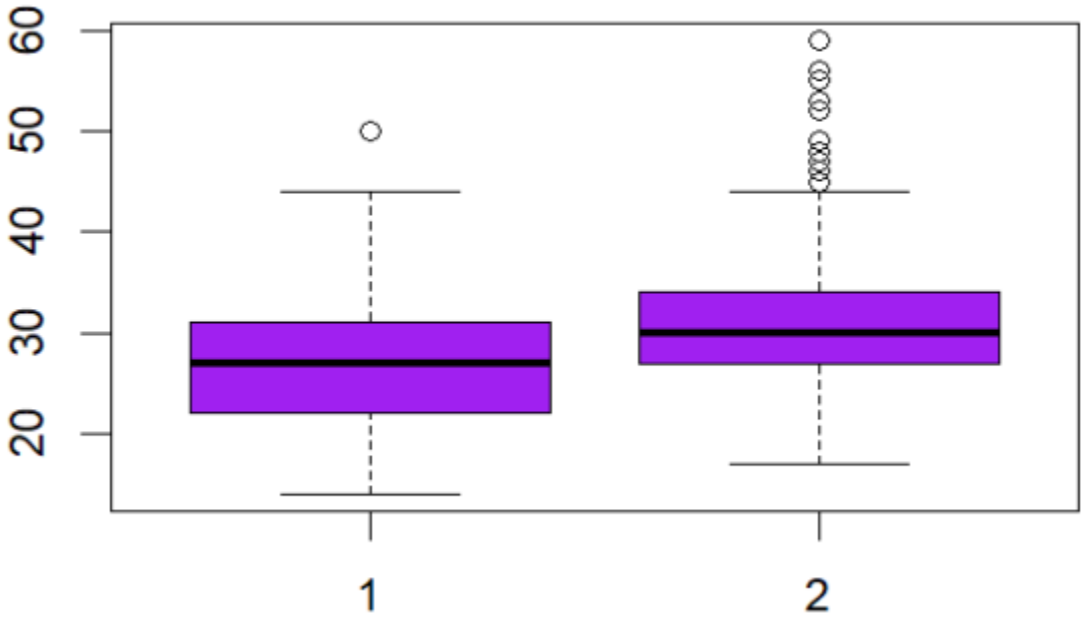

```
> histogram(weights_in_pounds, breaks=100, col="green")
```



The best visualization is given by the histogram with 20 bins as it clearly shows the shape of the graph and distribution of the data. It also doesn't disrupt the smoothness of the variations and can be understood and interpreted easily.

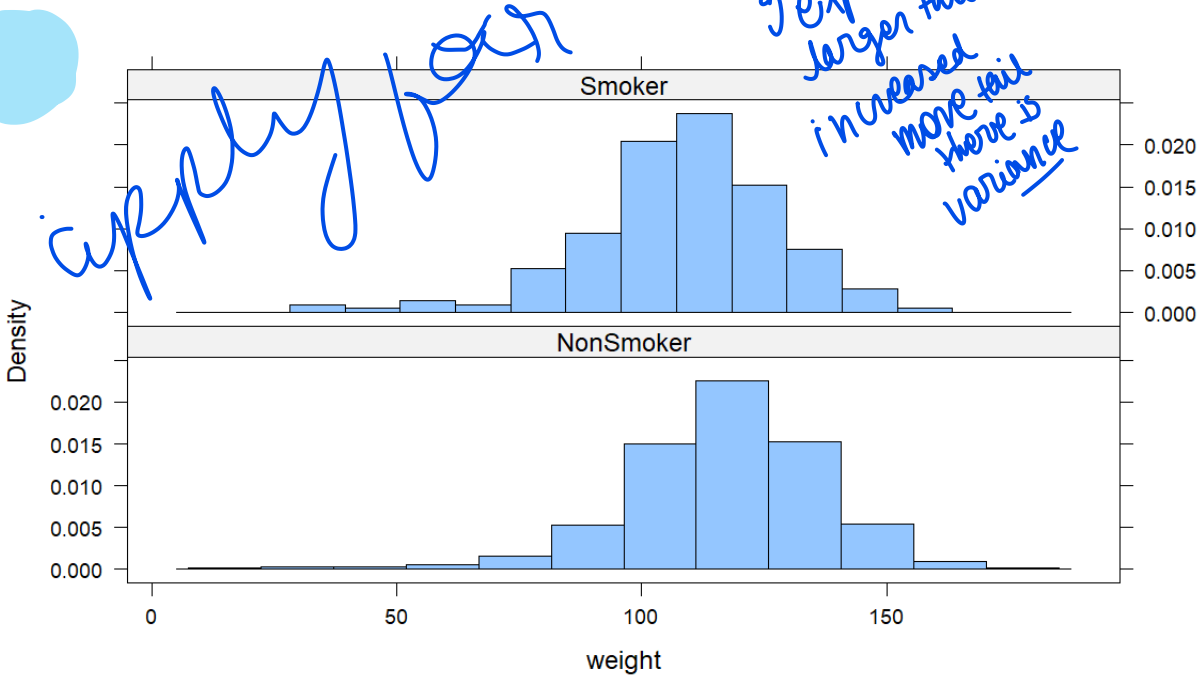
7. We can use the syntax boxplot(vector1, vector2) to make a side by side box plot. Create a side-by-side boxplot of the mother's ages and the father's ages. Which gender tends to be older?

```
> momage<-NCbirths$Mage
> fatherage<-NCbirths$Fage
> boxplot(momage, fatherage, col= "purple" )
```



The side by side box plot depicts that the father tends to be older than the mothers.

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?



The given code creates two stacked density vs. weight histograms corresponding to two categories in the habit variable. From the histograms, we can observe that the variability in weights of the babies of moms who smoke is higher than non-smoking moms.

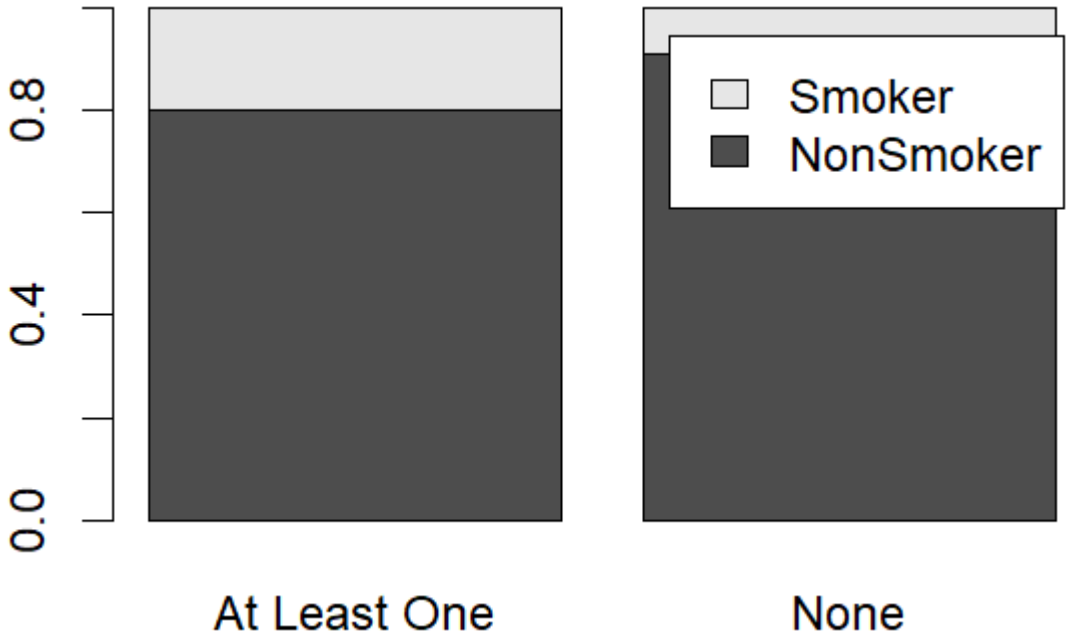
9. Produce a dot plot of the weights in pounds

```
>  dotPlot(weights_in_pounds, cex= 0.5, col= "red", xlab="Weights in pound
s", main="Dot plot of 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?

```
>  table1<-tally(~Habit | BirthDef, data = NCbirths, format = "proportion")
>  table1
           BirthDef
Habit       At Least One       None
  NonSmoker    0.80000000 0.90692969
  Smoker       0.20000000 0.09307031
```

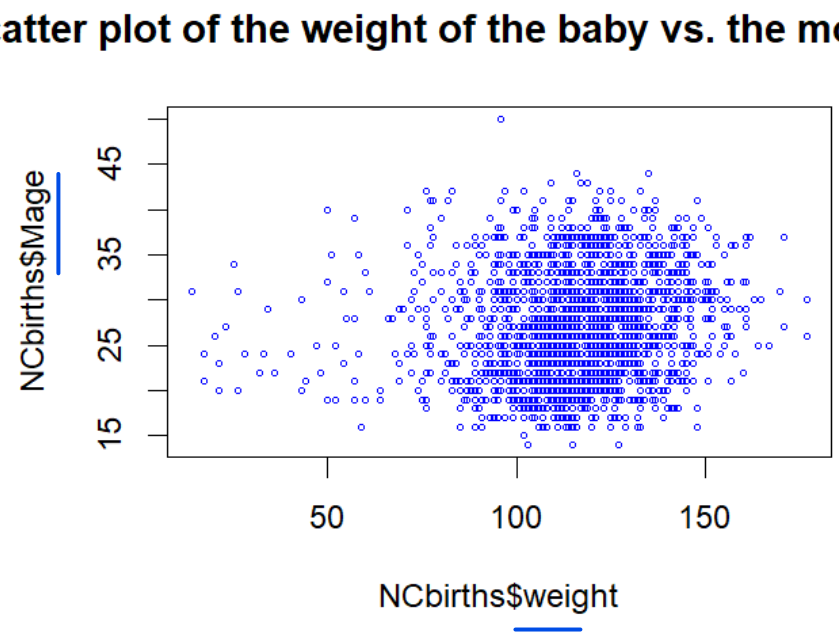
```
> table1<-tally(~Habit | BirthDef, data = NCbirths, format = "proportion")
> barplot(table1,legend.text=TRUE)
```



As smoking can affect the structure of DNA strands of the baby, birth defects are associated with smoking habit of the mother. As we can see in the two-way table, 90.692969% of children with no birth defects were conceived by a non-smoker mother, whereas a meagre 9.307031% of defect-free children had smoker moms. Due to a significant difference in the statistics, we can associate this variable with smoking.

11. Produce a nicely formatted scatter plot of the weight of the baby vs. the mother's age.

```
>  plot(NCbirths$weight, NCbirths$Mage, cex=0.5, col= "blue", main="Scatte
r plot of the weight of the baby vs. the mother's age")
```



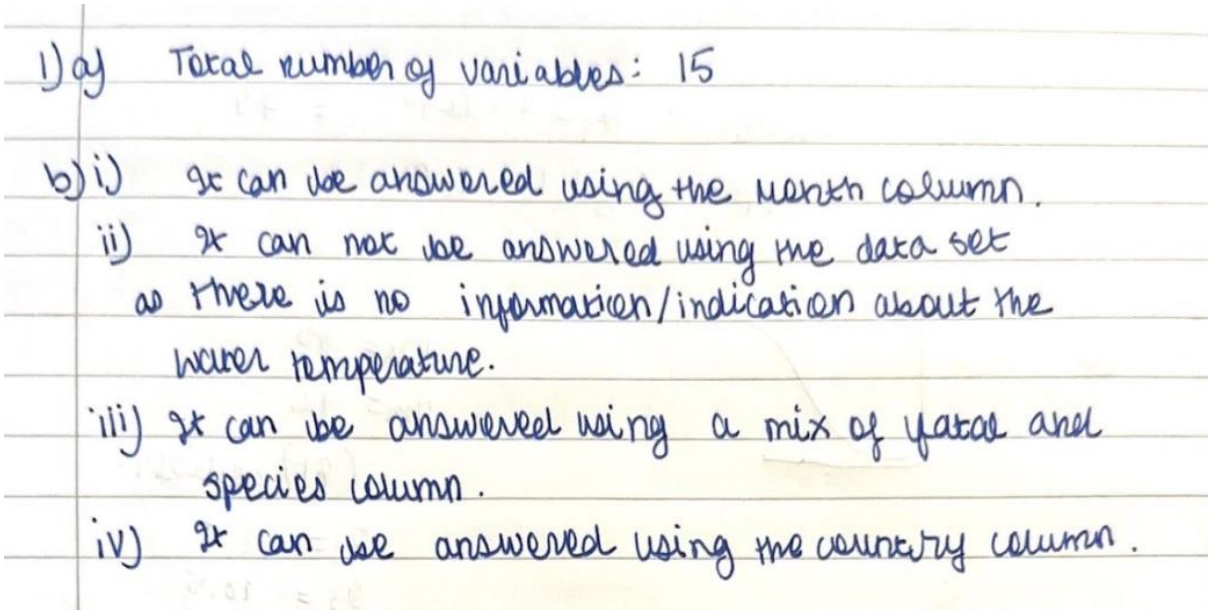
PART II

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

   a. How many variables are contained in the data?

   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?

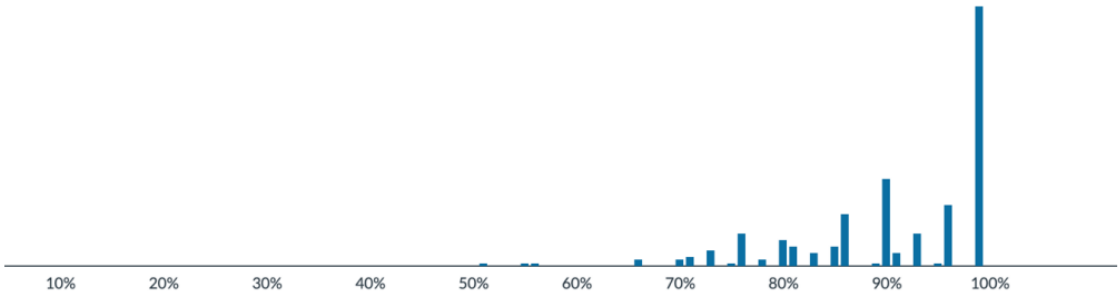
   c. A researcher wants to understand the age of the people in the data set and proposed some questions of interest: Are the reported cases are mostly younger people or older people? How is the age distributed? How would you help the research answer these questions? What statistical tools (e.g., graphs,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



c) We would first take a sample (subset) from the population (entire data) making sure that the observations we include in the sample have the ages mentioned. since some ages of some obs are missing. As it is a numerical data we will use a histogram to analyse it. Looking at the data, the most suitable bin width is of size 10.

To understand the age distribution we can look at the variability and the skewedness of the graph which can give us a hint about the mean, median and mode. We can then use the formula of standard deviation to measure the typical distance of the observation from the mean (measure the variability).

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



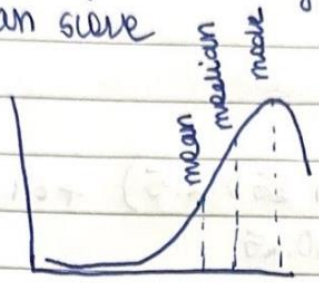
a. Describe the shape of distribution

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

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

2) a) since the tail is towards the left, it is a left skewed graph.

b) In a left skewed graph mean is less than the median score

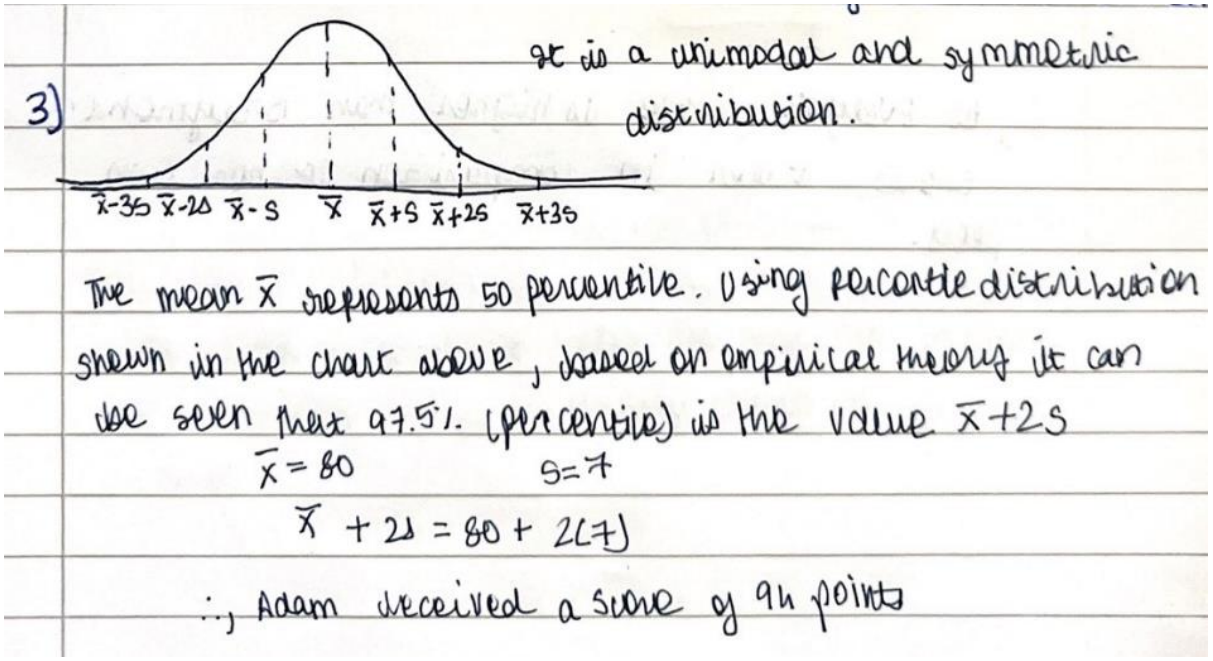


This is because the mean is affected more than the median by the large no. of smaller values.
The long tail of skewed graph and outliers affect the mean more than median.

c)

Median is a good measure of the center of a skewed distribution.
To measure the spread I would use the inter quartile range since it will be less influenced by the extreme low values. As it is the difference between Q_3 and Q_1 it measures the spread of data within the middle 50% of the distribution.

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.



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?

   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?

z score : $= \frac{x - \bar{x}}{s}$

4)

a) $x = ?$ $\bar{x} = 64$ $s = 2.5$ $z = -1.5$

$$z \text{ score} = \frac{x - 64}{2.5}$$

$$-1.5 = \frac{x - 64}{2.5}$$

$$x = (-1.25 \times 2.5) + 64$$

$$= 60.25$$

∴ the woman's weight corresponds to 60.25.

b) The magnitude of z score implies the unusualness of the observation.

Evelyn's zscore $= \frac{x - \bar{x}}{s} = \frac{75 - 64}{2.5}$

$= 4.40$

Draymond's z score $= \frac{x - \bar{x}}{s} = \frac{79 - 69}{3}$

$= 3.33$

∴ As Evelyn's z score is higher than Draymond's she is taller in comparison to her own 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.

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

5) Using the dataset given in the question

→ total no. of observation $(n) = 10$

a) minimum : 363

b) (Q_2) median : $\frac{\frac{n}{2}\text{th obs} + \frac{n}{2}+1 \text{ obs}}{2}$

$= \frac{5\text{th obs} + 6\text{th obs}}{2}$

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

minimum = 363 considering (obs below median): 363, 384, 389, 408, 423

$Q_1 = \frac{6\text{th obs}}{2} = 3\text{th obs} = 389$

considering obs above median: 434, 471, 520, 602, 677

$Q_3 = \frac{6\text{th obs}}{2} = 3\text{th obs} = 520$

maximum = 677

∴ the five number summary is:

min = 363, Q_1 = 389, Q_2 = 428.5, Q_3 = 520, max = 677

b) Using the five number summary we can calculate the range and the interquartile range which tells us about the distribution of the domestic gross income.

Range = max - min = 677 - 363 = 314

IQR = $Q_3 - Q_1$ = 131

We can also use the information to draw a box plot which tells us about the outliers.

If IQR is small it tells us that the income distribution is more tightly clustered around the median

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	Q1	Median	Q3	Maximum
1	62	91	218	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.
(The scale of the plot does not need to be accurate)

Not sure why I have lost points. Covered most of what the question requires

(a) min = 1

$Q_1 = 62$

(a) median = 91

$Q_3 = 218$

max = 756

$IQR = Q_3 - Q_1 = 156$

lower whisker = $Q_1 - 1.5(IQR) = -172$

upper whisker = $Q_3 + 1.5(IQR) = 452$

potential outliers: values that are greater than 452 are outliers.

outliers: 548 (Texas), 756 (New York)

The length of the whiskers are determined using the inter quartile range (IQR).

Usually,

the lower whisker boundary = $Q_1 - 1.5\,IQR$

the upper whisker boundary = $Q_3 + 1.5\,IQR$

However, since $Q_1 - 1.5\,IQR$ is lower than the min value in this case, the minimum value is the lower boundary.

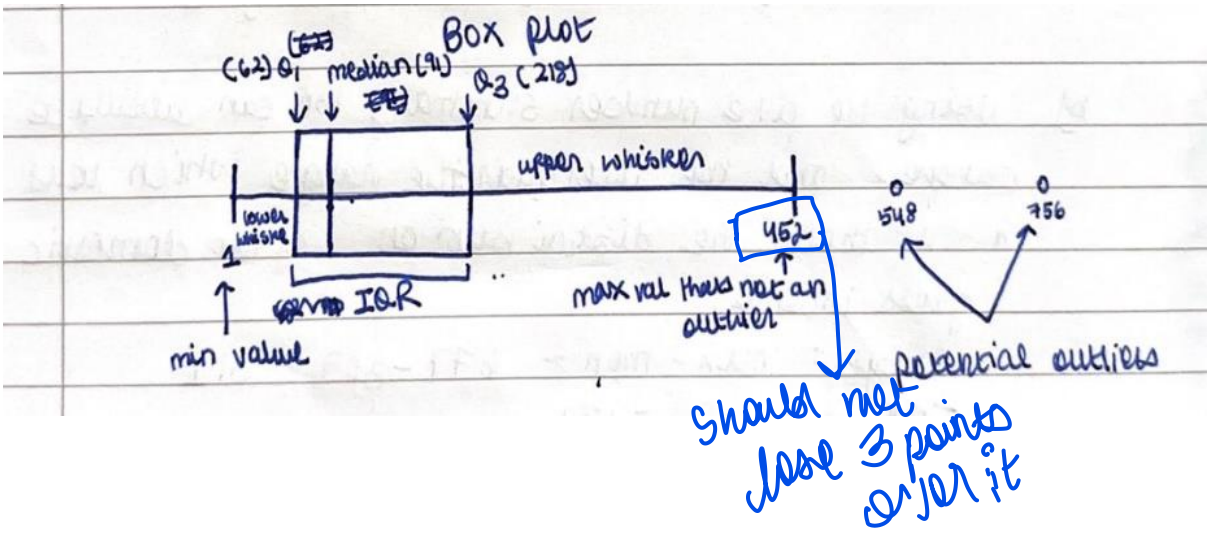
→ the length of lower whisker = Q_1 − min value

Since max value > $Q_3 + 1.5\,IQR$

The upper whisker boundary = $Q_3 + 1.5\,IQR$

→ The length of the upper whisker = $Q_3 + 1.5\,IQR - Q_3$

$= 1.5\,IQR$.


Box plot
(62) Q1
median (94)
Q3 (218)
upper whisker
lower whisker
1
IQR
452
max val that is not an outlier
min value
548
756
potential outliers
should not close 3 points or for it