**Part I**

1. a)

**Command:**

A white background with black text

Description automatically generated

**Output:**

A close-up of a number

Description automatically generated

b)

**Command:**

A close-up of a person

Description automatically generated

**Output:**

A close-up of words

Description automatically generated

c)

**Output:**

A black and white text with black text

Description automatically generated

🡪 The *cbind(heights, names)* function combines the two vectors *heights* and *names* into a data frame where each column represents a variable. The first column contains heights in inches and the second column contains names of the students.

A close-up of a text

Description automatically generated

🡪 This new object cbind(heights, names) is the matrix type of an array.

1. a), b)

Command:

A close-up of a computer screen

Description automatically generated

**Output:**

> NCbirths <- read.csv("births.csv")

> 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

1. a)

**Command:**

A close-up of a website

Description automatically generated

**Output:**

A screenshot of a computer

Description automatically generated

b)

**Output:**

A screenshot of a computer

Description automatically generated

1. a), b), c), d)

**Command:**

A computer screen shot of a number of text

Description automatically generated with medium confidence

🡪 The weights of the babies should be in the units ‘grams’.

**Output:**

A number on a white background

Description automatically generated

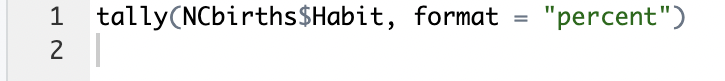
1. Mean weight of the babies in pounds:

**Output:**

A close-up of a computer code

Description automatically generated

**Command:**



**Output:**

A white background with black text and numbers

Description automatically generated

🡪 Therefore, approximately 9.38% of the mothers in the sample smoke.

1. Comparing the percentage obtained from the NCbirths data set to the CDC’s report, the percentage is about 11.62% off.

**Command:**

A white background with black text

Description automatically generated

**3 bins:**

A graph of weight loss

Description automatically generated

20 bins:

A graph of weight loss

Description automatically generated

100 bins:

A graph of weights and weights

Description automatically generated

🡪 The histogram with *20 bins* seems to give the best visualization since it strikes a balance between the 3 bins which shows large ranges of weight in each bin thereby hiding finer details, and the 100 bins which shows too much detail and noise thereby making it hard to identify patterns. Thus, the histogram with 20 bins provides a good balance between capturing the underlying data distribution and maintaining a clear and interpretable visualization.

**Command:**

**A close-up of a sign

Description automatically generated**

**Output:**

A graph with lines and dots

Description automatically generated

🡪 It can be deduced that the men tend to be older since the position of the boxplot for *father’s ages* as well as the median is positioned higher than the boxplot *for mother’s age* and its median, indicating that there are more men who are older in age.

**Output:**

**A graph of a graph of smoker and non smoker

Description automatically generated**

🡪 This code creates a side-by-side histogram of baby weights based on whether the mother smokes or not (basically grouping the weight variable by the habit variable). It can be interpreted as:

1. **~ weight | Habit:** This part specifies the formula for the histogram and indicates that we want to create histograms of the "weight" variable, and the data should be grouped by the "Habit" variable (smoking or non-smoking moms).
2. **data = NCbirths:** This argument specifies the dataset, which is "NCbirths" in this case.
3. **layout = c(1, 2):** This argument specifies the layout of the histograms and arranges them in one row and two columns, creating side-by-side plots.

Based on the graph, it can be observed that both graphs have the same centre, are unimodal in the number of peaks and have little variability or spread and therefore, there aren’t any major differences between baby weights from smoking moms vs. non-smoking moms.

**Command:**

dotchart(weights\_in\_pounds)

**Output:**

**A graph showing a number of dots

Description automatically generated**

🡪 Of the categorical variables in this data that record the health of the baby, the ones that are associated with the mother’s smoking are:

*Premie* (premature birth), *Apgar1* (score for appearance, pulse, grimace, activity, and respiration), *BirthDef* (birth defects), and *BirthComp* (birth complications), since all these are measures that reflect the conditions and state of the baby at birth, and are thus correlated to the mother’s smoking habits, which is a major factor affecting them.

**Command:**

A close up of a text

Description automatically generated

**Output:**

A white background with black text

Description automatically generated

🡪 The evidence that the variable *Premie* (premature birth) is associated with maternal smoking is indicated by the differences in proportions of maternal smoking habits (*Habit*) between babies with and without premature birth in the two-way Summary Table:

**Proportion of Smokers in Babies Without Premature Birth (No Premie)**:

* Among mothers of babies without premature birth ("No Premie"), approximately 9.11% are smokers ("Smokers").
* In contrast, the majority, about 90.89%, are non-smokers ("NonSmokers").

**Proportion of Smokers in Babies With Premature Birth (Yes Premie)**:

* Among mothers of babies with premature birth ("Yes Premie"), there is a higher proportion of smokers. Approximately 12.15% are smokers ("Smokers").
* About 87.85% are non-smokers ("NonSmokers").

**Key Observations:**

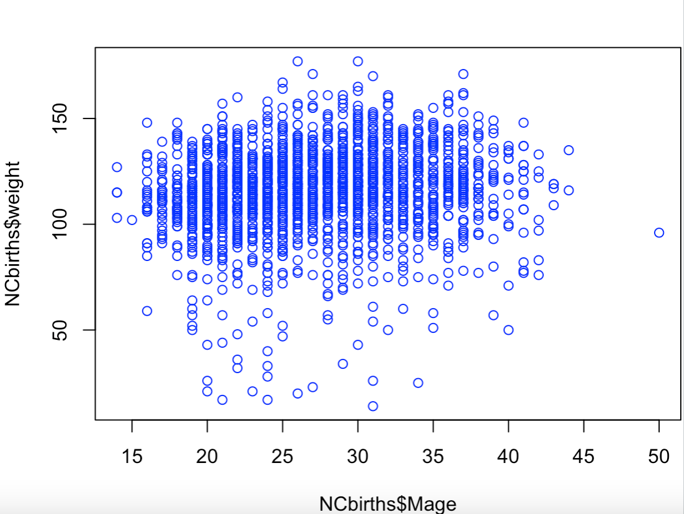
The proportions of smokers among mothers of babies with premature birth are higher compared to mothers of babies without premature birth. This suggests that there is an association between maternal smoking and the likelihood of premature birth.

11.

**Command:**

plot(NCbirths$weight ~ NCbirths$Mage, col = "blue", cex = 1)

**Output:**

****

**Part II**

1. 15 variables
2. This can be answered since the ‘Month’ column can be analysed to tally in which month the most shark attacks occur.
3. This can be answered since temperature-related variables like ‘Area’ and ‘Location’ can be analysed and tallied to observe whether shark attacks more likely to occur in warm temperature or cooler temperatures.
4. This can be answered by using the ‘Species’ and ‘Fatal (y/n)’ variables in a two-way summary table to analyse which shark species are associated with a higher likelihood of fatalities in shark attacks.
5. This can be answered using the ‘Country’ variable to identify the country with the highest number of shark attacks per year.
6. The statistical tools the researcher can use are:
7. Summary Tables- Provide age statistics for specific categories to compare age distributions among subgroups.
8. Histograms- Visualize the entire age distribution, revealing its shape and identifying common age ranges.
9. Box Plots- Illustrate age distribution spread, central tendencies, and identify outliers and thus useful for comparing age distributions between different groups.
10. The shape of the distribution is multimodal since it has multiple peaks.
11. The mean score would be less than the median score since the graph observed is left skewed in nature and thus, the long tail of the skewed distribution typically pulls the mean (average) to the left of the median (middle value).
12. The measures used to report the centre and spread would be:

🡪 *Centre Measure*- Median, since it is less affected by outliers and skewed data, making it a robust measure of the central tendency. In a left-skewed distribution, the median is closer to the typical (central) score.

🡪 *Spread Measure-* Interquartile range (IQR) instead of the standard deviation, since the IQR is calculated using the 25th and 75th percentiles (the lower and upper quartiles) and is less sensitive to extreme values, which is important in skewed distributions and thus provides a robust measure of spread.

1. Can use the Z-score formula:

In this case, we assume that the distribution of test scores in the class is approximately normally distributed, given that it's described as unimodal and symmetric.

**Formula: Z = (x - x̄)/ s**

Z is the Z-score

x is the value we want to find (Adam's score)

x̄ is the mean (**80** points)

s is the standard deviation (**7** points)

Since the distribution is symmetric, we can use the standard normal distribution table. Using the table, we find that a percentile of 97.5% corresponds to a z-score of approximately **1.96.**

Solving for x to find Adam’s score:

x = (1.96 x 7) + 80

**x = 93.72**

So, Adam received a score of approximately **93.72 points**. This score is **higher** than approximately 97.5% of the students in the class, as he estimated.

1. Using the Z score formula:

**Z = (x - x̄)/ s**

Z is the Z score (-1.50)

x is the value we want to find (women’s height)

x̄ is the mean height for women (64 inches)

s is the standard deviation (2.5 inches)

Solving for x to find women’s height:

x = (-1.50 x 2.5) + 64

**x = 60.25**

Therefore, a Z-score of -1.50 corresponds to a women’s height of **approximately 60.25 inches.**

For Evelyn Akhator:

Height (x) = 75 inches

Mean height for women (x̄) = 64 inches

Standard deviation for women’s heights (s) = 2.5 inches

**Z-score of Evelyn = (75 – 64)/2.5 = 4.4**

For Draymond Green:

Height (x) = 79 inches

Mean height for men (x̄) = 69 inches

Standard deviation for men’s heights (s) = 3 inches

**Z-score of Draymond = (79 – 69)/3 = 3.33**

Comparing the Z-scores,

Evelyn Akhator's Z-score within the women's group is 4.4, indicating she is significantly taller than her peers.Draymond Green's Z-score within the men's group is 3.33, indicating he is also significantly taller than his peers. Therefore, both players are taller than their respective peers, but in a direct comparison, Evelyn Akhator has a higher Z-score within her gender group, meaning she is relatively taller compared to her peers than Draymond Green is compared to his peers.

2. Five-number summary:

**Minimum (Min):** 363 million

**First Quartile (Q1):** 389 million

**Median (Q2):** 428.5 million

**Third Quartile (Q3):** 520 million

**Maximum (Max):** 677 million

1. Interpretation of the five-number summary:

* The minimum domestic gross income of any of the top ten Marvel comic-based movies is 363 million (from "Deadpool").
* The median domestic gross income is 428.5 million, which means that half of the movies in the top ten list have gross incomes below this value, and half have gross incomes above it.
* The first quartile (Q1) is around 389 million, suggesting that the lowest 25% of the movies have domestic gross incomes less than this value.
* The third quartile (Q3) is approximately 520 million, indicating that the lowest 75% of the movies have domestic gross incomes less than this value.
* The maximum gross income is 677 million (from "The Avengers"), which is the highest among the top ten Marvel comic-based movies.

Overall, this five-number summary provides information about the range and distribution of domestic gross income for these movies, showing that there is considerable variability in box office success among them.

Lower whisker: Starts from the minimum (1)

Lower edge of the box: Q1 (62)

Middle line of the box: Median (91)

Upper edge of the box: Q3 (218)

Upper whisker’s end point:

IQR = Q3 - Q1 = 218 - 62 = 156

1.5 \* IQR = 234

whisker can extend up to Q3 + 234 = 452. However, this doesn't reach the maximum (756). The data point 756 is considered an outlier since it's beyond 1.5 times the IQR from Q3.