DATA CLEANING
Xinyi (Anny) Cui

Before we start deal with our data file, we need to modify the downloaded file in Excel to make sure it can be read in R. We need to convert the txt file to csv file by copy and paste all data from the downloaded into Excel and using "Text to Columns" button to adjust each column and then save it as csv file.

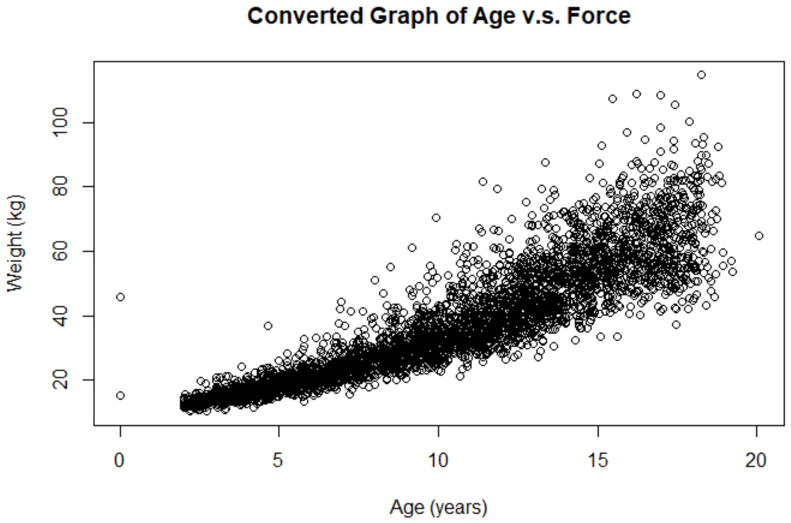
1. There are 123 variables in the data file.

2. There are 3901 observations in the data file.

3. There are 2600 children do not have a recorded head length.

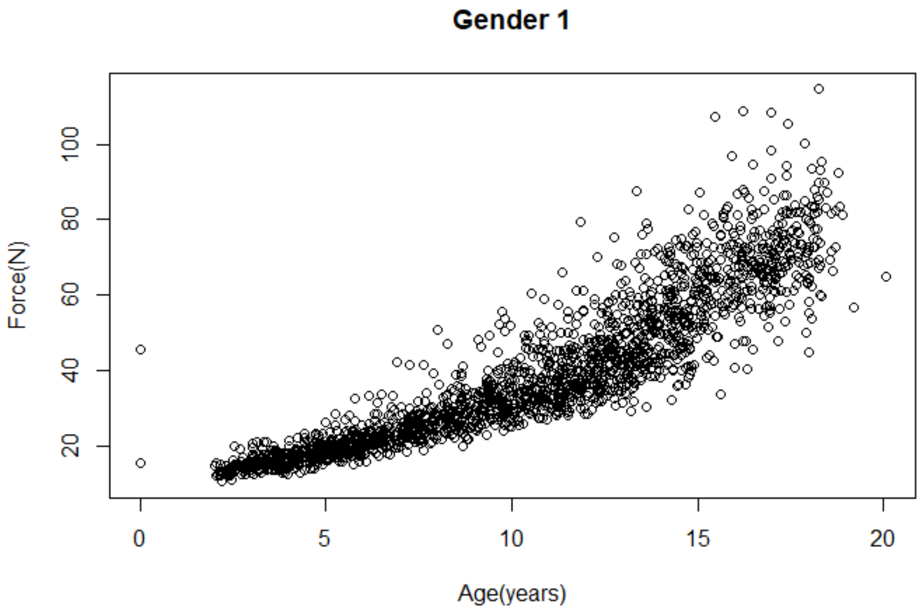
4. We can infer the units of age in years are $\frac{1}{1000}$ years. From the final report of ANTHROPOMETRY OF INFANTS,CHILDREN, AND YOUTHS TO AGE 18 FOR PRODUCT SAFETY DESIGN on May 31, 1977, we learn that the study randomly selects children whose ages are 0 to 18 years old (pp.1-9). While, when we check the column of Age in Years, we find that the ages are in thousands or in over ten thousands. After checking the range of this column, we can see that the smallest number in this column is 0 and the largest one is 20054. Seems like multiply the year age by 1000. Also, this report is an observation from 1975 to 1977 ("over the 22-month period of this study" (pp. i), which last for almost 2 years. Therefore, we can infer that 20054 is a child who participate in 1975 at the age of 18.0054, and this make sense on most weird numbers for age in years.

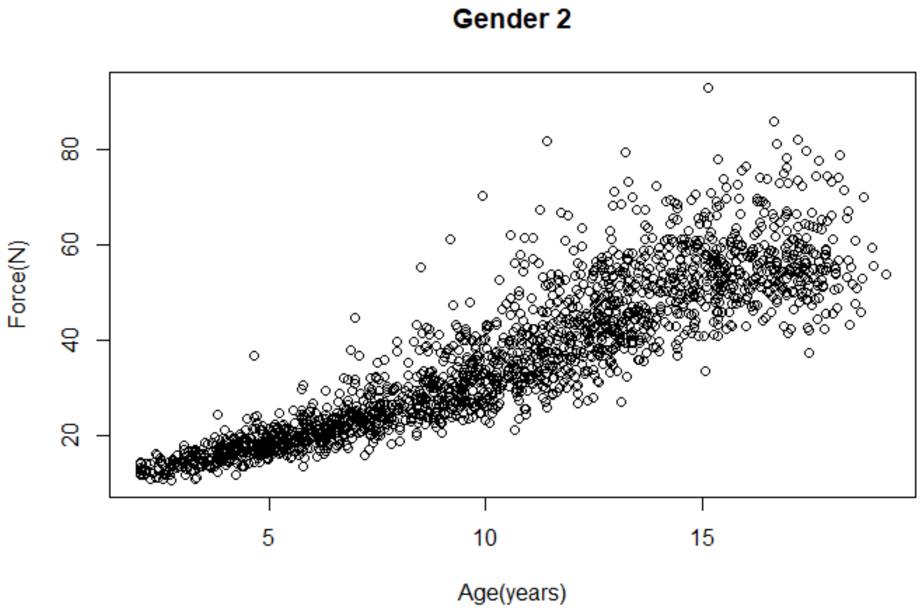
5. On page of 43 in the report, we can see that "all results in this report are given in metric units To convert kilograms (kg) to pounds (lbs) multiply by 2.2". We can also check the graph of weight (kg) against the age (years) on page of 61 in the report. Individually, after comparing the original graph to the graph shown in the report, we can assume the units for weight are newtons ($1kg = 9.8N$).



Therefore, we can infer that the better name for this column is Force.

6. We can use graphs on page of 62 in the report to compare with plots in our data and then see 1 and 2 represent male or female separately. The plots are shown below:


Gender 1


Gender 2

Then, We can assume that gender 1 represents males and gender 2 represents females as those graphs are like each other corresponding plots in the reports.

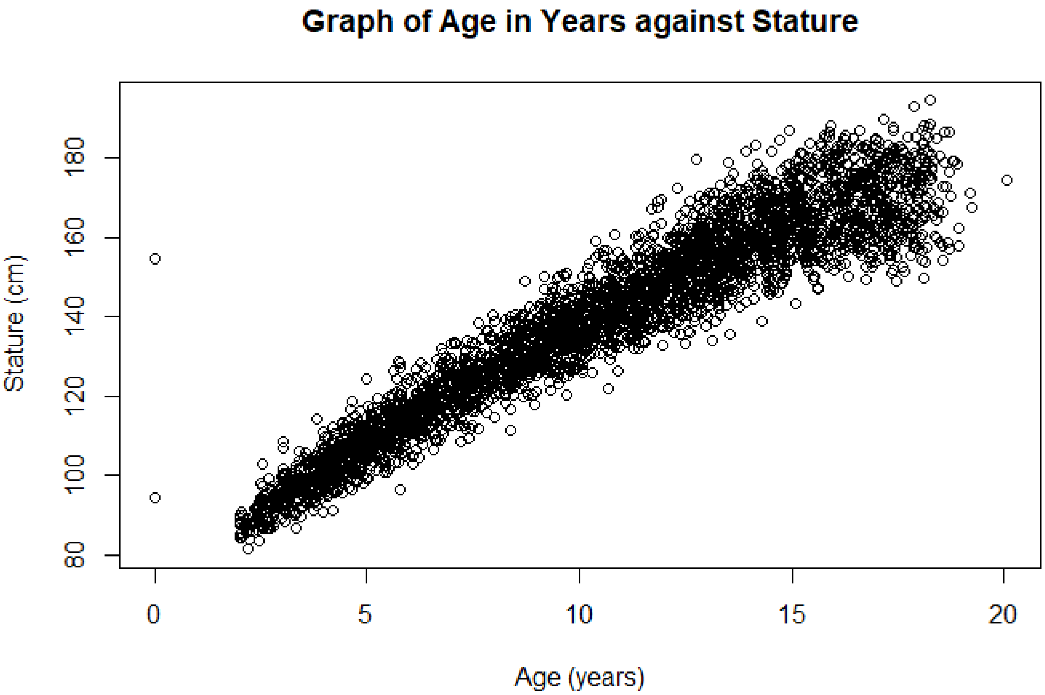
7-9 We can use our intuition to analyze problem 7 to 9. (I literally find no clues in the report or data file for those problems.)

7. There are 3240 children have handedness equals to 1 and 352 kids have handedness equals to 2. From our daily life, we can assume that number 1 means right-handed and number 2 represents left-handed. That is because the rarity of left-handed persons in our daily life.

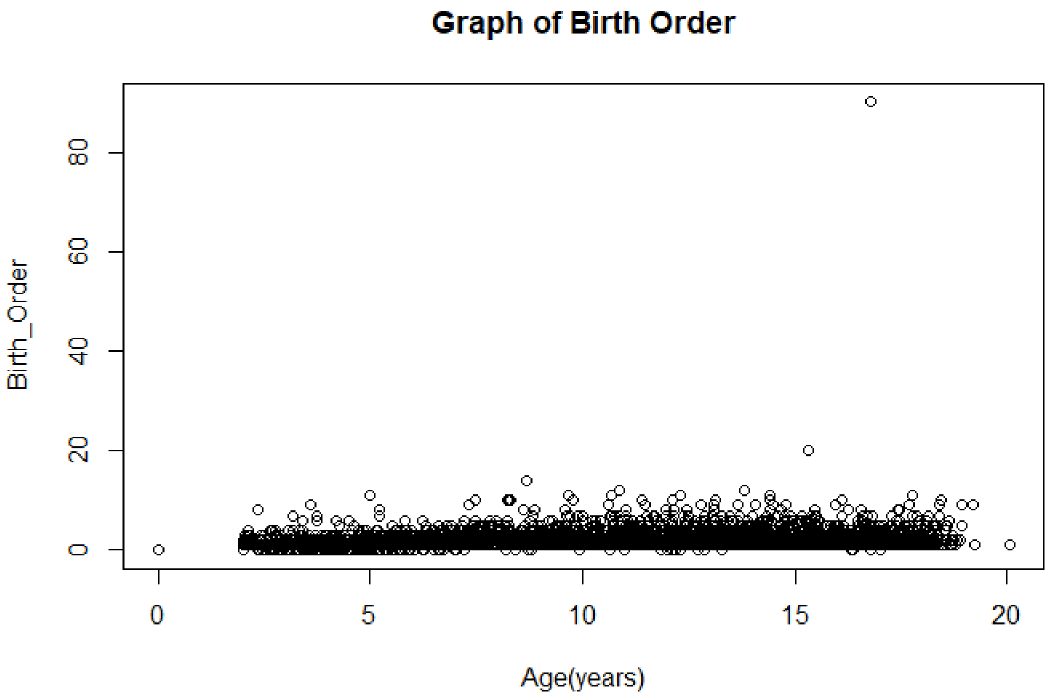
8. There are 309 children whose "Handedness" listed as a number other than 1 or 2.

9. Besides 1 and 2, number 0 (115 kids), 3 (193 kids), 4 (0 kids), and 5 (1 kid) are used to record for handedness. The other numbers might mean that kids are good at using both hands, unknown (because some are infants who do not know which hand they prefer to use), unsure (do not know if they are left or right handed), or none (disability).

10. The stature is the vertical distance from the standing surface to vertex. "subject stands erect with head oriented in the Frankfort Plane, arms hanging at sides. With an automated anthropometer, measure the vertical distance from the standing surface to vertex (top of the head)" (pp. 64). In data file, the statures are recorded in millimeters (mm). Therefore, if we divide the values of stature by 10 to convert them into centimeters, the graph seems match with the one on page of 65 in the report, shown below:

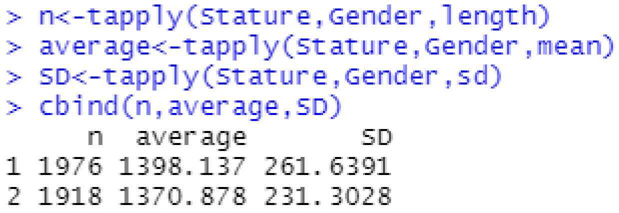


11. On page of 16 in the report, in section 4(g), we can see that the birth order represents this child's relationship to his or her siblings (1st oldest, 2nd oldest, etc.). After graphing, we can see some points which are obviously outliers. The graph is shown below:

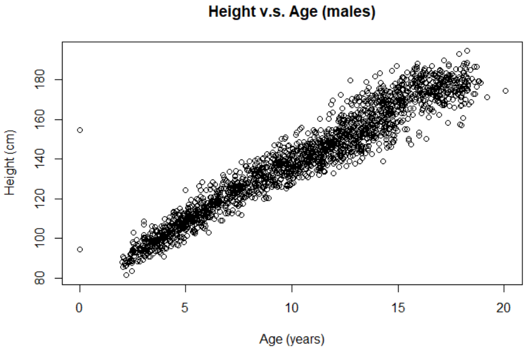


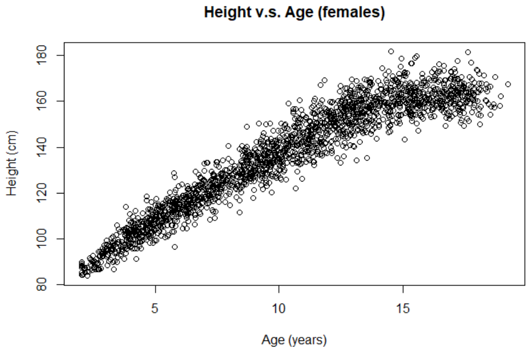
Obviously, there are at least 2 points unusual, which are the one over 20 and another one over 90. These points probably represent 2nd oldest and 9th oldest, but they are recorded with one more zero after the integers by mistake.

12. This question asks to compare the heights of male and female. There is a bunch of different heights and the measurement is not clarified, so I assume it asks the stature comparison.



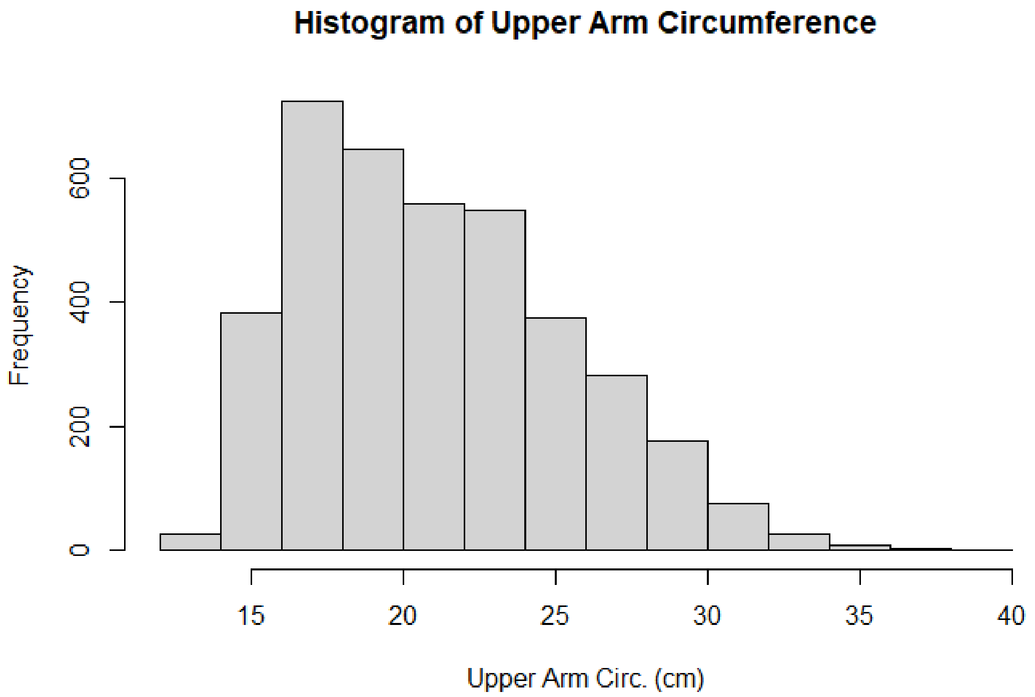
```
> n<-tapply(Stature,Gender,length)
> average<-tapply(Stature,Gender,mean)
> SD<-tapply(Stature,Gender,sd)
> cbind(n,average,SD)
     n  average       SD
1 1976 1398.137 261.6391
2 1918 1370.878 231.3028
```



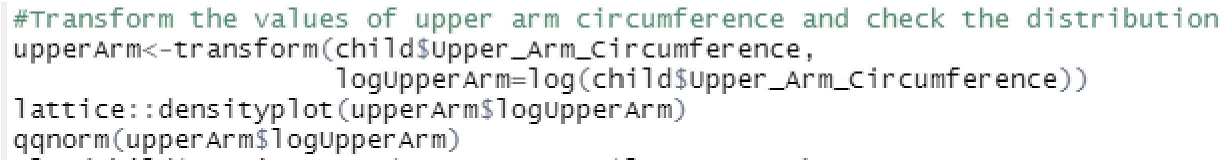


Linear Regression Models and their summary are shown in R file.

13. The histogram of upper arm circumference is left-skewed shown below:

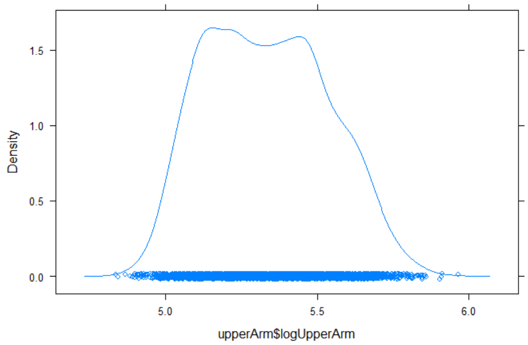


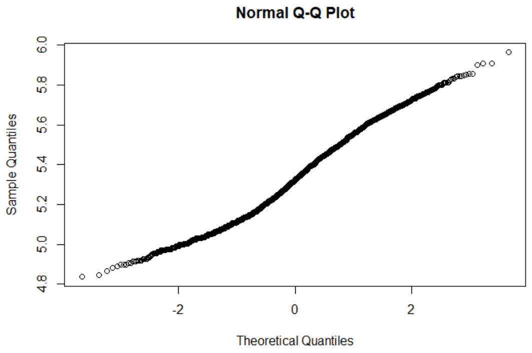
Using the functions shown below to transform values of upper arm circumference:



```
#Transform the values of upper arm circumference and check the distribution
upperArm<-transform(child$Upper_Arm_Circumference,
                         logUpperArm=log(child$Upper_Arm_Circumference))
lattice::densityplot(upperArm$logUpperArm)
qqnorm(upperArm$logUpperArm)
```

Then, we can check with the density plot and the normal quantile graph:





The distribution looks more like a bell-shaped normal distribution with less skewness, and the qqnorm graph displays a more linear line which means the values have a more linear relationship.