```
# SS 2022
# Descriptive Statistics
# a) import the file galtonfamilies.csv as a tibble called galtonfamilies
library(readr)
galtonfamilies <- read_csv("D:/Datentransfer/Studium/3.Semester/Statistics/
SS_22/GaltonFamilies.csv")
View(GaltonFamilies)
tb <- read.csv("D:/Datentransfer/Studium/3.Semester/Statistics/SS_22/
GaltonFamilies.csv")
tb
str(tb)
# b) Determine the scale of all variables. You find a description of the
dataset
#   in the file GaltonFamilies_Description.pdf

#+  family(id) : qualitative discrete nominal
#+  father(height): quantitative, continuous ratio
#+  mother(height):  quantitative, continuous ratio
#+  midparentHeight:  quantitative, continuous ratio
#+  children: quantitative , discrete , absolute   possible value 0 ??
#+  childNum: qualitative, discrete, ordinal
#+  gender: qualitative discrete nominal
#+  childHeight:  quantitative, continuous  ratio

# c) The height is given in inches. Change the values to cm ( 1 inch =
2.54cm)
galtonfamilies <- galtonfamilies %>%
  mutate(father = father*2.54,
         mother = mother*2.54,
         midparentHeight = midparentHeight*2.54,
         childHeight = childHeight*2.54)

galtonfamilies

# d) Create a tibble heights.fm containing the heights of the fahers and
#    mothers in the families. The tibble should have the two columns
#    "type" and "Height". The variable "type" indicates if the value
#    of height bleongs to a father or mother

heights.fm <- galtonfamilies %>% gather('father', 'mother', key = type,
value = height) %>%
  select(type,height)
heights.fm

# e) Create a summary describing the distribution of the variable height in
#    the dataset heights.fm containing n,min,max,mean,median,Q1,Q2,Q3,
#    depeneding on the variable type

measures.fm <- heights.fm %>% group_by(type) %>%
          summarise(
            n = n(),
            min = min(height),
            max = max(height),
            mean = mean(height),
            median = median(height),
            Q1 = quantile(height, 0.25, type=1),
            Q2 = quantile(height, 0.5, type=1),
            Q3 = quantile(height, 0.75, type=1),
            iqr = IQR(height, type=1)
          )
measures.fm
# f) create a side by side boxplot for the height of persons depending
#    on their sex and interpret the diagram.
boxplot(heights.fm$height~heights.fm$type)
# both of the boxplots are rights skewed because the median is near
# to the first quantile.
# there are 4 extreme values in the group of fathers, which means there
# are some fathers with really low height(3 fathers) and one father whom
# height larger
# the min of the fathers is higher than the females
# the max of the fathers is higher than the females


# i) The file children.csv contains the data of 50 additional children.
#    The heights of the parents are given in the file parents.csv.
#    Import both files to the tibbles children and parents.

parents<- read_csv("D:/Datentransfer/Studium/3.Semester/Statistics/SS_22/
parent.csv")
children <- read_csv("D:/Datentransfer/Studium/3.Semester/Statistics/SS_22/
children.csv")
children
parents
# j) Complete the missing height of the parents in the tibble children
#    and add the data to the dataset galtonfamilies.
# left join
galtonfamilies
children <- children %>% left_join(parents, by='...1')
children
# full_join
galtonfamilies %>% full_join(children, by = '...1') %>% select(-X.x,-X.y)
```