```r
options(repos = c(CRAN = "https://cran.rstudio.com/"))
install.packages('dafs')
```

```
## package 'dafs' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\zheng\AppData\Local\Temp\RtmpKmyYno\downloaded_packages
```

```r
library(dafs)
```

```
## Warning: package 'dafs' was built under R version 4.3.3
```

```
## Loading required package: s20x
```

```
## Warning: package 's20x' was built under R version 4.3.3
```

```r
data(dpd.df)
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.3.3
```

First of all, explore the data and check for any missing values: There is no any missing value in dataset, so we can start.

```r
View(dpd.df)
summary(dpd.df)
```

```
##      sample           age        sex        molar         dpd.ratio
##  Min.   : 1.00   Min.   :15.00   F:13   Min.   :16.00   Min.   :0.620
##  1st Qu.: 6.25   1st Qu.:21.75   M: 9   1st Qu.:27.00   1st Qu.:1.087
##  Median :11.50   Median :38.50          Median :37.00   Median :1.265
##  Mean   :11.50   Mean   :41.23          Mean   :36.09   Mean   :1.427
##  3rd Qu.:16.75   3rd Qu.:60.00          3rd Qu.:46.75   3rd Qu.:1.587
##  Max.   :22.00   Max.   :73.00          Max.   :48.00   Max.   :3.050
##     est.age       assoc.error      real.error
##  Min.   :24.00   Min.   :14.60   Min.   :-29.7000
##  1st Qu.:33.88   1st Qu.:14.60   1st Qu.: -7.2250
##  Median :37.70   Median :14.70   Median :  0.1500
##  Mean   :41.15   Mean   :14.91   Mean   :  0.8045
##  3rd Qu.:44.60   3rd Qu.:14.88   3rd Qu.:  8.4500
##  Max.   :75.70   Max.   :16.80   Max.   : 26.0000
```

```r
dim(dpd.df)
```

```
## [1] 22  8
```

```r
sum(is.na(dpd.df))
```

```
## [1] 0
```

```r
colSums(is.na(dpd.df))
```

```
##      sample         age         sex       molar   dpd.ratio     est.age
##           0           0           0           0           0           0
## assoc.error  real.error
##           0           0
```

Q1. How many male individuals are there in this study? How many female individuals are there in this study?

```
summary(dpd.df$sex)
```

```
##  F  M
## 13  9
```

```
print(paste("Male individuals are there in this study is:", paste(table(dpd.df$sex)['M'])))
```

```
## [1] "Male individuals are there in this study is: 9"
```

```
print(paste("Female individuals are there in this study is:", paste(table(dpd.df$sex)['F'] )))
```

```
## [1] "Female individuals are there in this study is: 13"
```

Q2. What is the mean and median age of the female individuals in this study?

```
Fe_data <- subset(dpd.df, sex == "F")
Ma_data <- subset(dpd.df, sex == "M")

mean_Fe_age <- mean(Fe_data$age, na.rm = TRUE)
median_Fe_age <- median(Fe_data$age, na.rm = TRUE)


print(paste("The mean and median age of the female individuals in this study are:", paste(mean_Fe_age),paste("and",paste(median_Fe_age))))
```

```
## [1] "The mean and median age of the female individuals in this study are: 40.3076923076923 and 37"
```

Q3. What is the range for the age of the male individuals in this study? What is the range for the age of the female individuals in this study?

```
range_Ma_age <- range(Ma_data$age,na.rm = TRUE)
range_Fe_age <- range(Fe_data$age,na.rm = TRUE)

print(paste("The female age range is:", paste(range_Fe_age, collapse = " - ")))
```

```
## [1] "The female age range is: 15 - 73"
```

```
print(paste("The male age range is:", paste(range_Ma_age, collapse = " - ")))
```

```
## [1] "The male age range is: 17 - 61"
```

Q4.What is the maximum DPD ratio for the male individuals in this study? What is the maximum DPD ratio for the female individuals in this study?

```
max_Ma_DPD_ratio <- max(Ma_data$dpd.ratio,na.rm = TRUE)
max_Fe_DPD_ratio <- max(Fe_data$dpd.ratio,na.rm = TRUE)

print(paste("The maximum DPD ratio for the male individuals is:", paste(max_M
a_DPD_ratio)))
```

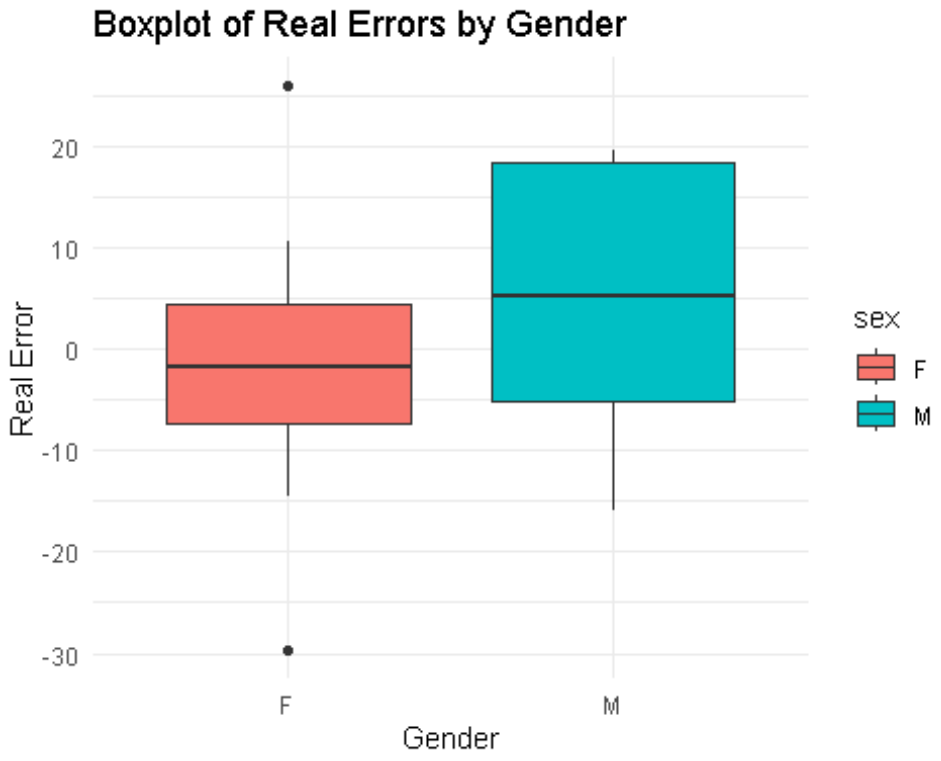
```
## [1] "The maximum DPD ratio for the male individuals is: 1.69"
```

```
print(paste("the maximum DPD ratio for the female individuals:", paste(max_Fe
_DPD_ratio)))
```

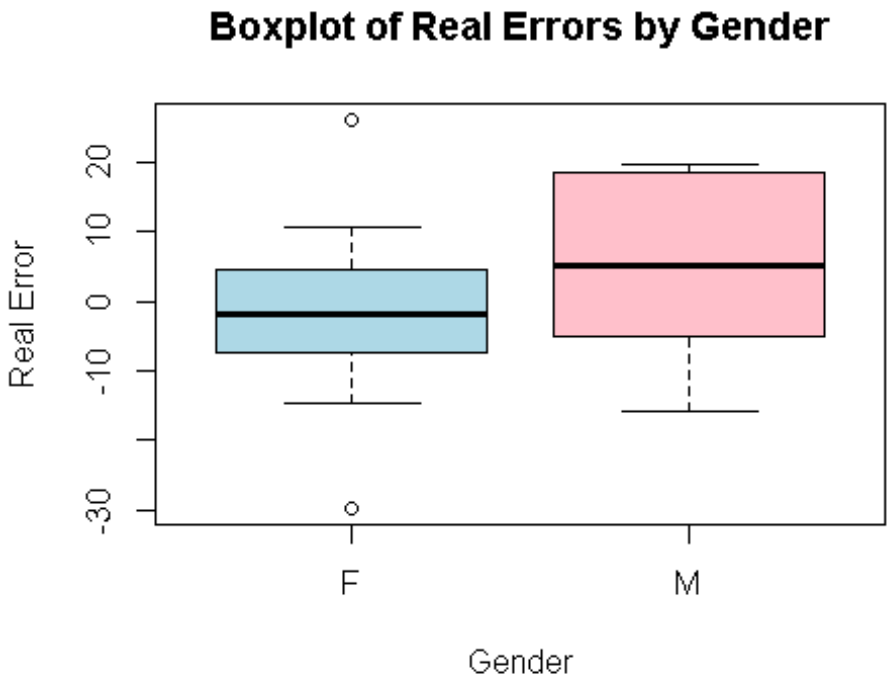
```
## [1] "the maximum DPD ratio for the female individuals: 3.05"
```

Q5.To see whether there is a difference between the actual errors for the female individuals and the actual errors for the male individuals, create a box and whisker plot comparing the actual errors (i.e.,"real.error") obtained for female individuals with the actual errors (i.e., "real.error") obtained for male individuals (i.e., create one graph containing two box and whisker plots, one for females and one for males). Interpret your answer.

```
ggplot(dpd.df, aes(x = sex, y = real.error, fill = sex)) +
  geom_boxplot() +
  labs(title = "Boxplot of Real Errors by Gender", x = "Gender", y = "Real Er
ror") +
  theme_minimal()
```



```r
boxplot(real.error ~ sex, data = dpd.df,
        main = "Boxplot of Real Errors by Gender",
        xlab = "Gender",
        ylab = "Real Error",
        col = c("lightblue", "pink"))
```



Interpretation of the Boxplot Comparing Real Errors for Male (M) and Female (F) Individuals:

1. Median: The median for male individuals (M) is higher than for female individuals (F), indicating that males tend to have higher real errors compared to females.
2. IQR: The IQR for male individuals is wider than that for female individuals, meaning that there is more variability in the real errors among males than females.
3. Whiskers: For the range within 1.5 times the IQR, males extend -16 to 19, female -14 to 19.
4. Outliers: No outlier for the male but there are 2 outliers for females. One is higher than 25 and the other one is around -30.

Q6.What is the interquartile range for the actual error for female individuals? What is the interquartile range for the actual error for male individuals?

```r
IQR_Ma <- IQR(Ma_data$real.error,na.rm = TRUE)
IQR_Fe <- IQR(Fe_data$real.error,na.rm = TRUE)
```

```
print(paste("Interquartile Range for Female Individuals:", IQR_Fe))
```

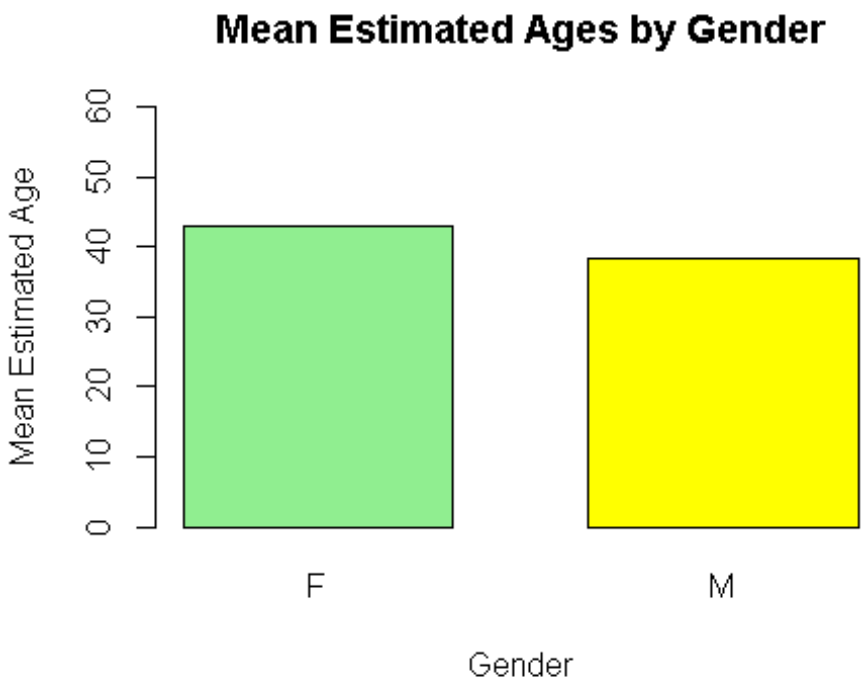
```
## [1] "Interquartile Range for Female Individuals: 11.9"
```

```
print(paste("Interquartile Range for Male Individuals:", IQR_Ma))
```

```
## [1] "Interquartile Range for Male Individuals: 23.6"
```

Q7.To compare the estimated ages for each gender, create a bar plot showing the mean estimated ages for each gender.

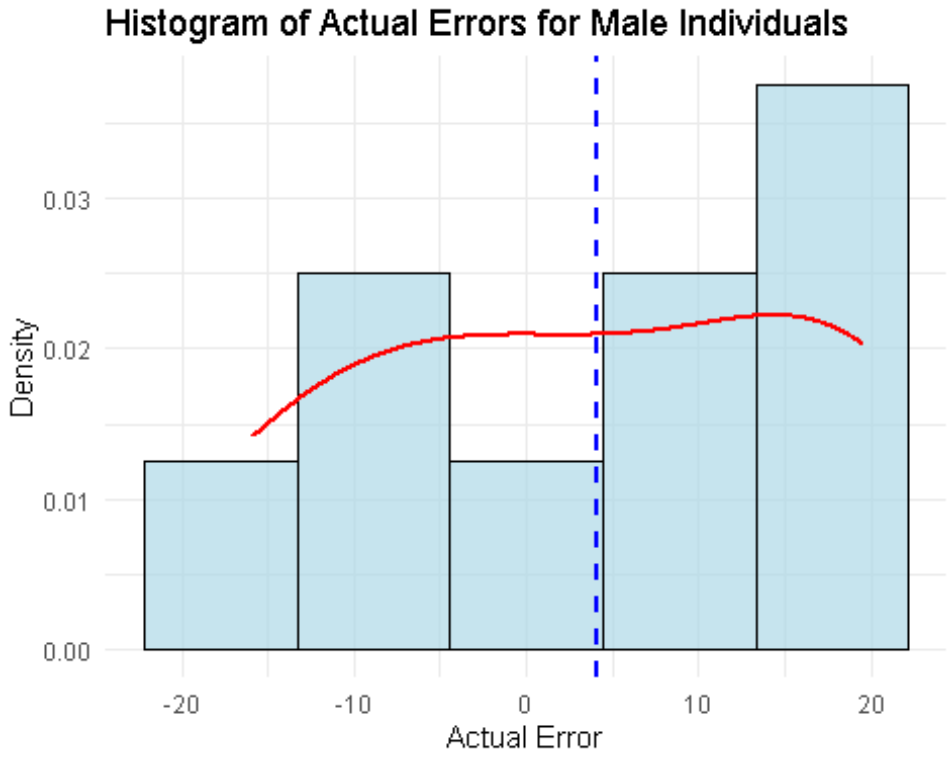
```
mean_est_age <- aggregate(est.age ~ sex, data = dpd.df, FUN = mean)
barplot(mean_est_age$est.age,
        names.arg = mean_est_age$sex,
        col = c("lightgreen", "yellow"),
        main = "Mean Estimated Ages by Gender",
        xlab = "Gender",
        ylab = "Mean Estimated Age",
        width = 0.5,
        space = 0.5,
        ylim = c(0, 60))
```



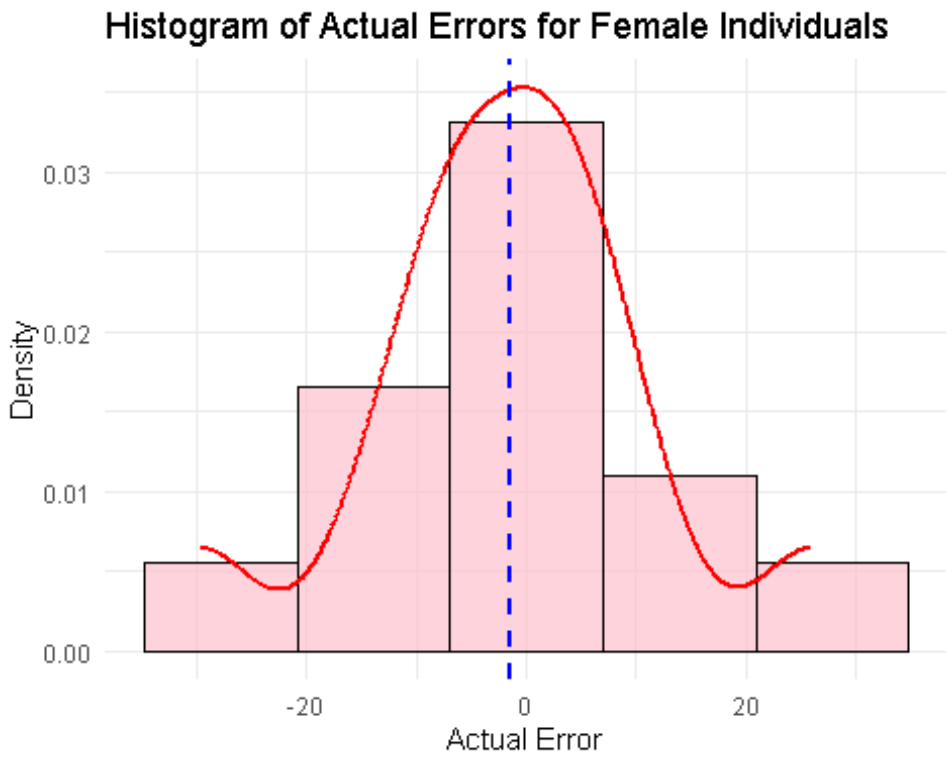
Q8. Represent the actual errors for the female individuals and the actual errors for the male individuals as separate histograms. In each histogram, plot a line representing the density and a vertical line marking the mean value of the actual errors. Inter pret your answer.

```r
mean_Ma_re.er <-mean(Ma_data$real.error, na.rm = TRUE)
mean_Fe_re.er <-mean(Fe_data$real.error, na.rm = TRUE)

ggplot(Ma_data, aes(x = real.error)) +
    geom_histogram(aes(y = after_stat(density)), bins = 5, fill = "lightblue"
, color = "black", alpha = 0.7) +
    geom_density(color = "red", linewidth = 1) +
    geom_vline(aes(xintercept = mean_Ma_re.er), color = "blue", linetype = "d
ashed", linewidth = 1) +
    labs(title = "Histogram of Actual Errors for Male Individuals",
         x = "Actual Error",
         y = "Density") +
    theme_minimal()
```




```r
ggplot(Fe_data, aes(x = real.error)) +
    geom_histogram(aes(y = after_stat(density)), bins = 5, fill = "pink", col
or = "black", alpha = 0.7) +
    geom_density(color = "red", linewidth = 1) +
    geom_vline(aes(xintercept = mean_Fe_re.er), color = "blue", linetype = "d
ashed", linewidth = 1) +
    labs(title = "Histogram of Actual Errors for Female Individuals",
         x = "Actual Error",
         y = "Density") +
    theme_minimal()
```



Histogram of Actual Errors for Male Individuals

1. The histogram shows that the actual errors for male individuals are not symmetrically distributed.
2. The actual errors range from approximately -20 to +20.
3. The red density curve does not show a distinct peak.
4. The blue dashed line marks the mean of the actual errors around 4.

Histogram of Actual Errors for Female Individuals

1. The histogram and the red density curve suggest that the actual errors for female individuals are distributed symmetrically around zero.
2. Unlike the male error distribution, this plot shows a distinct peak in the density curve at zero.
3. Most of the errors fall within the range of -20 to +20.
4. The blue dashed line indicates the mean error, which is very close to zero.