Exploratory Data Analysis (EDA)

Visual Summaries

It is a key part of what we do when we analyze data. We start out every analysis with EDA to familiarize ourselves with the data. Before applying any technique, we use several visual plot or summary statistics to make sure the data are in agreement with the necessary assumptions.

Histograms Histograms are, generally used, to summarise numerical data. Here, each observation will be added into an interval in order to spot which values are the most (or less) frequent in your dataset. Through this visual plot, you will have an idea about the distribution of your data by looking at the shape of the histogram.

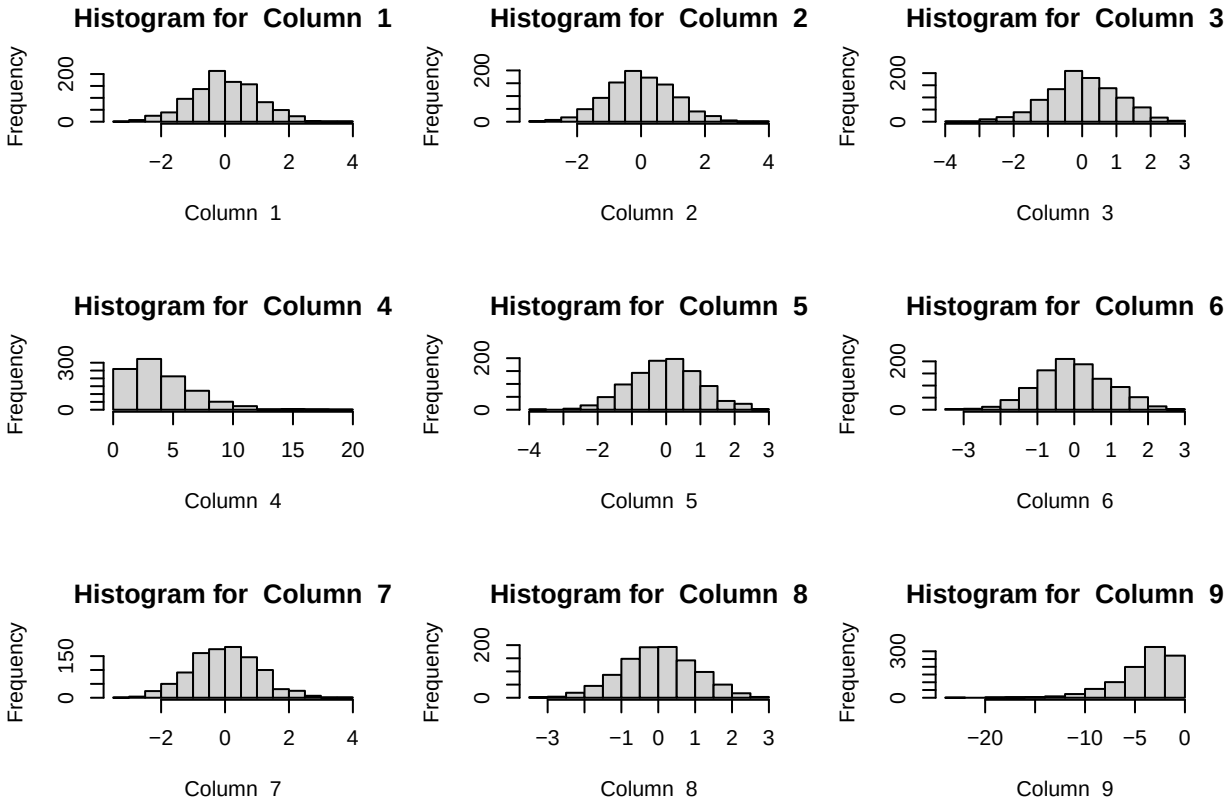
```
library(readr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

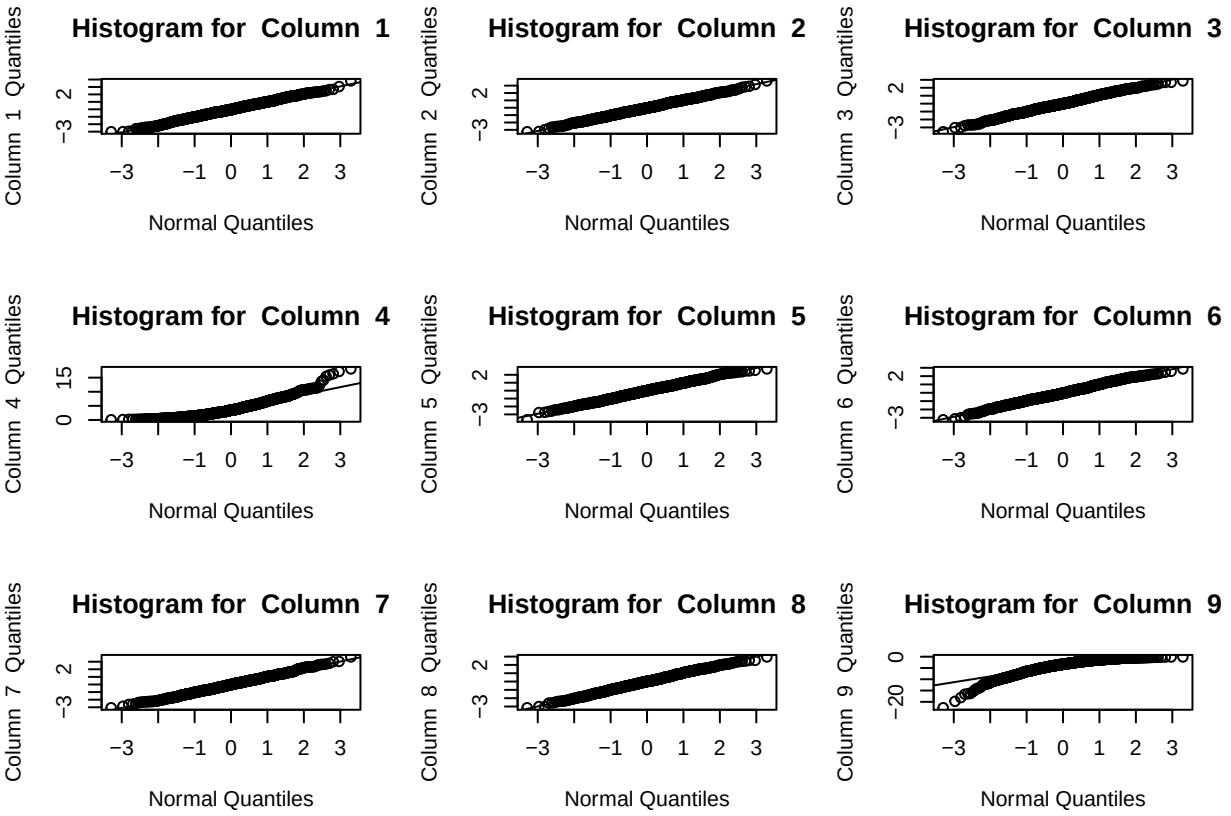
```
load("data/skew.RData")
par(mfrow = c(3,3))
for (i in 1:9) {
  column_name <- paste("Column ",i)
  main_title <- paste("Histogram for ",column_name)
  hist(dat[,i], main=main_title, xlab=column_name)
}
```



As we can see from the histogram shapes, most of the columns seem normally distributed except for the 4th and 9th columns. The 4th column is right-skewed (or has a positive skew) since it shows a long tail to the right (toward larger values). The 9th column is left-skewed (or has a negative skew) since it shows a long tail to the left (toward smaller values).

Quantile-Quantile plots (Q-Q plot) This visual plot will help you find out which distribution is best suited for your data. We assume the data follow a known distribution, we will use the normal distribution here. We generate quantiles for our data and for our Normal distribution.

```
par(mfrow = c(3,3))
for (i in 1:9) {
  column_name <- paste("Column ", toString(i))
  ylabel <- paste(column_name, " Quantiles")

  main_title <- paste("Histogram for ",column_name)
  qqnorm(dat[,i],main=main_title, xlab="Normal Quantiles",ylab=ylabel)
  qqline(dat[,i])
}
```



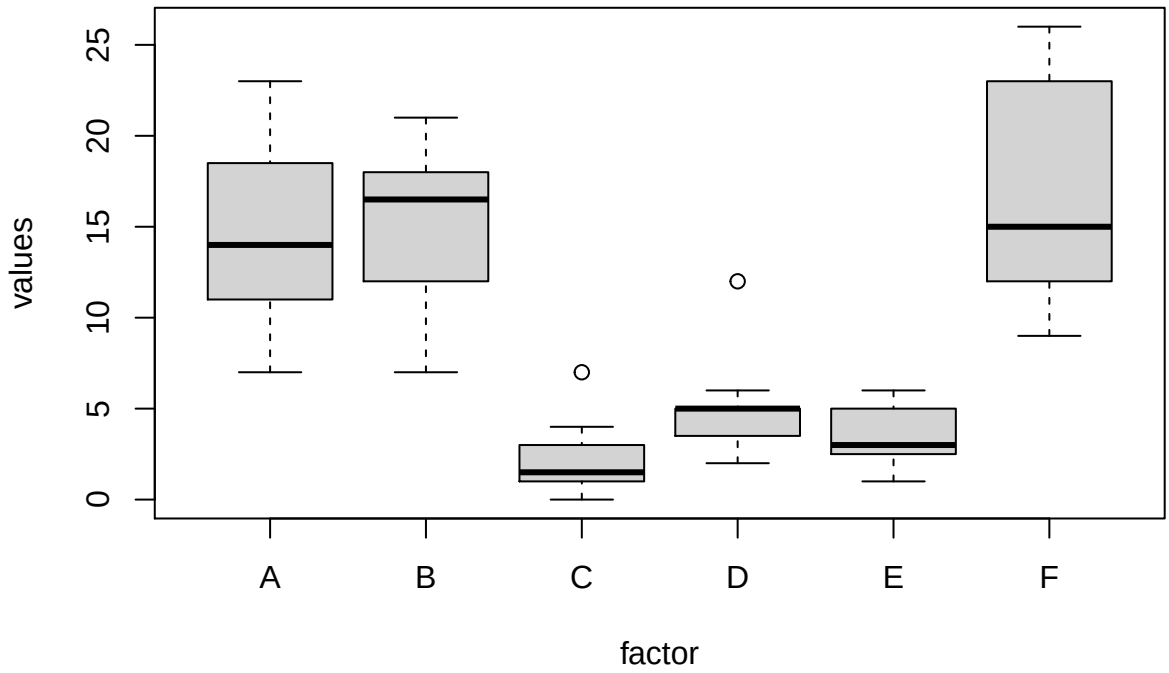
If the data follow the Normal distribution, then most of the points on the q-q plot will fall approximately on the identity line, this is what we can see for most of the columns. Note that positive skew looks like an up-shaping curve just like the Q-Q plot for the 4th column, while negative skew looks like a down-shaping curve just like the Q-Q plot for the 9th column. The 4th and 9th columns are then the only columns that are not normally distributed.

Boxplots When the data is not normally distributed, another interesting way of graphically summarizing data is boxplots. It gives you a summary regarding five numbers which are: the median, the 25th and 75th quantiles, the min and max values (which give you the range of values of your data).

50% of the data is above the median while the other remaining 50% is below the median. Within the box itself, we have 25% of the data above the median and 25% of the data below the median. Outliers are also visible as dots in the graph.

We will use the InsectSprays data set which measures the counts of insects in agricultural experimental units treated with different insecticides. This dataset is included in R, and you can examine it by typing:

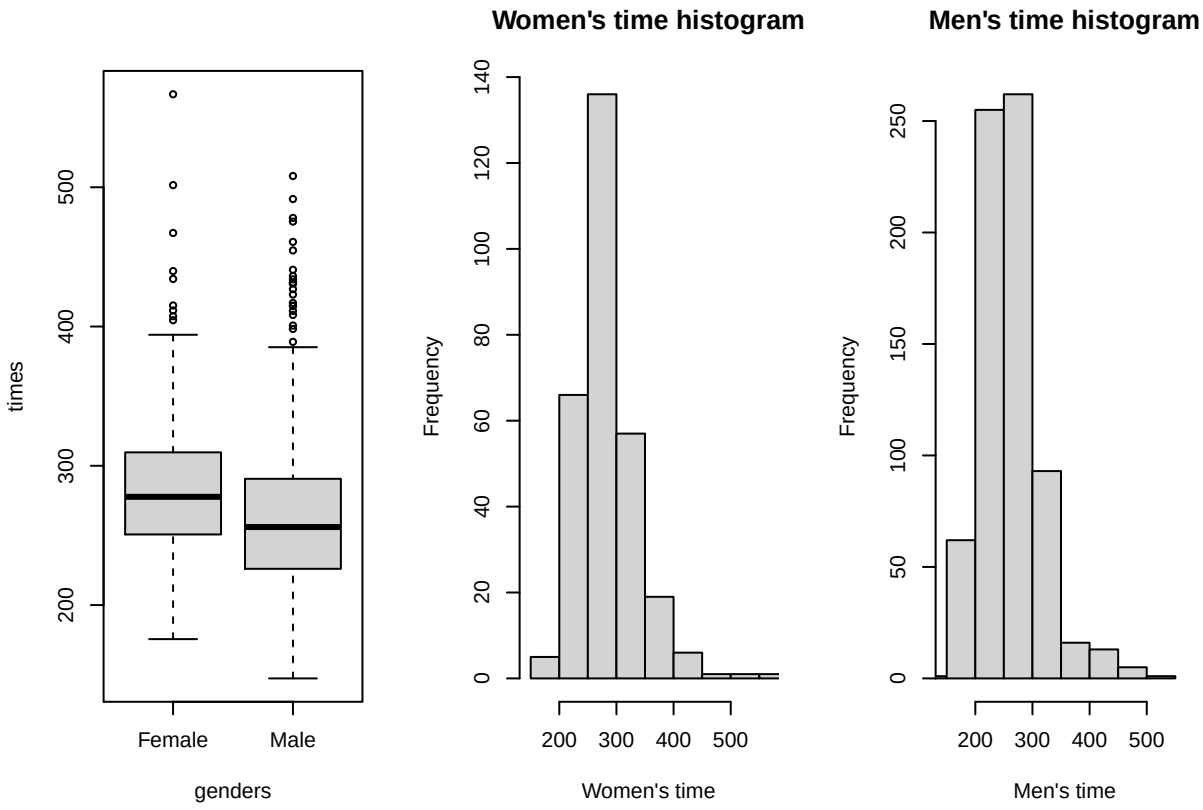
```
values <- InsectSprays$count
factor <- InsectSprays$spray
boxplot(values ~ factor)
```



We will use boxplots to compare the effectiveness of different sprays. With this boxplot, we can spot the most effective spray which is the spray having the lowest median count. Thus,it looks like the most effective one is the C spray.

To go further, let's consider a random sample of finishers from the New York City Marathon in 2002. This dataset can be found in the UsingR package. We will compare Male and Female finishing times.

```
par(mfrow = c(1,3))

#Boxplots
data(nym.2002, package = "UsingR")
genders <- nym.2002$gender
times <- nym.2002$time
boxplot(times~genders)

#Histograms

women_times <- filter(nym.2002, gender=="Female") %>% select(time) %>% unlist
men_times <- filter(nym.2002, gender=="Male") %>% select(time) %>% unlist
hist(women_times, xlim=c(range( nym.2002$time)),main="Women's time histogram", xlab="Women's time")
hist(men_times,xlim=c(range( nym.2002$time)),main="Men's time histogram",xlab="Men's time")
```




```
cat('Average time for Men:',mean(men_times),', ','Average time for Women:',mean(women_times))
```

```
## Average time for Men: 261.8209 ,  Average time for Women: 284.9366
```

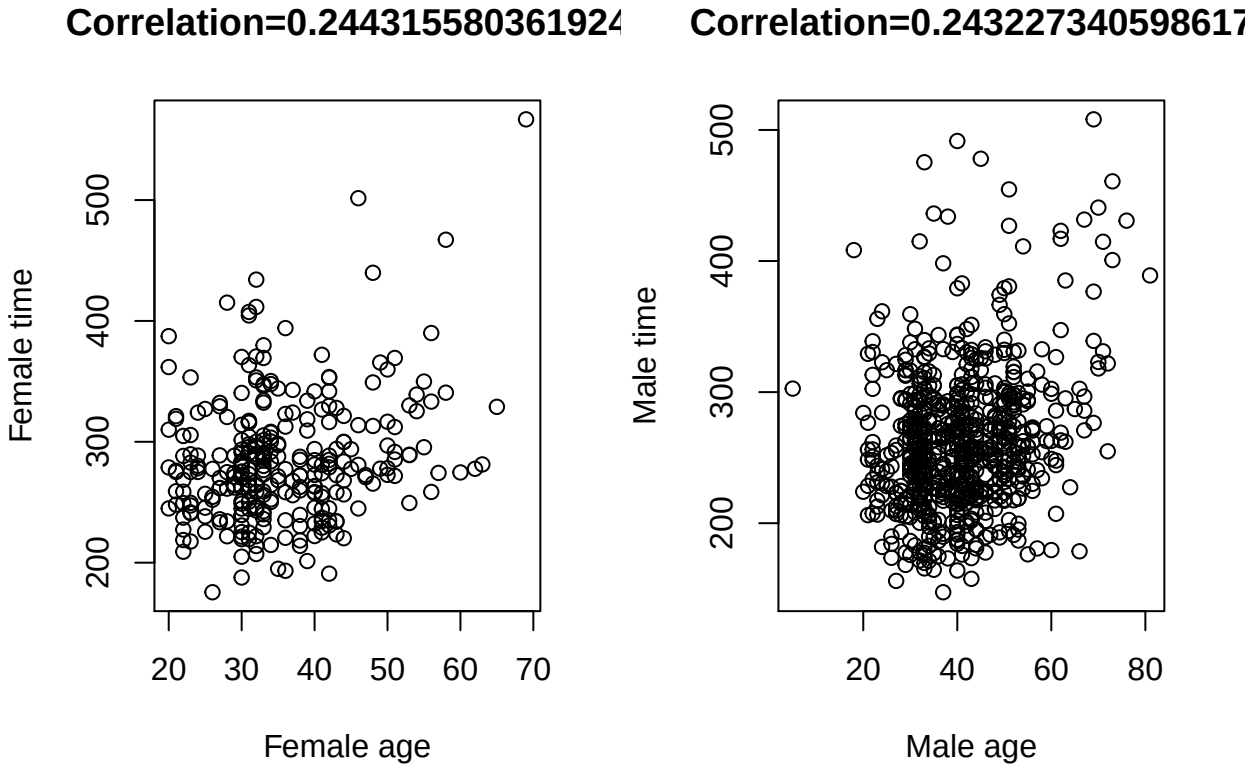
```
cat("Means difference between Women and Men:",mean(women_times) - mean(men_times))
```

```
## Means difference between Women and Men: 23.11574
```

Males and females have similar right skewed distributions with the former, 20 minutes shifted to the left.

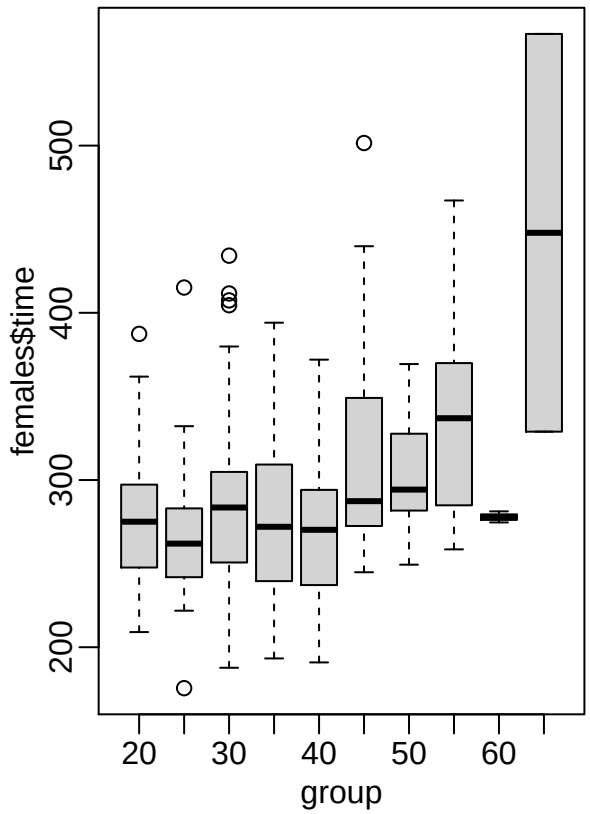
Scatter plots 2D Scatter plots are another way of visualizing continuous data and spot if there is any linear relationship between 2 variables. The linear relationship can be quantified by the correlation value telling us how strong the relation between both variables is.

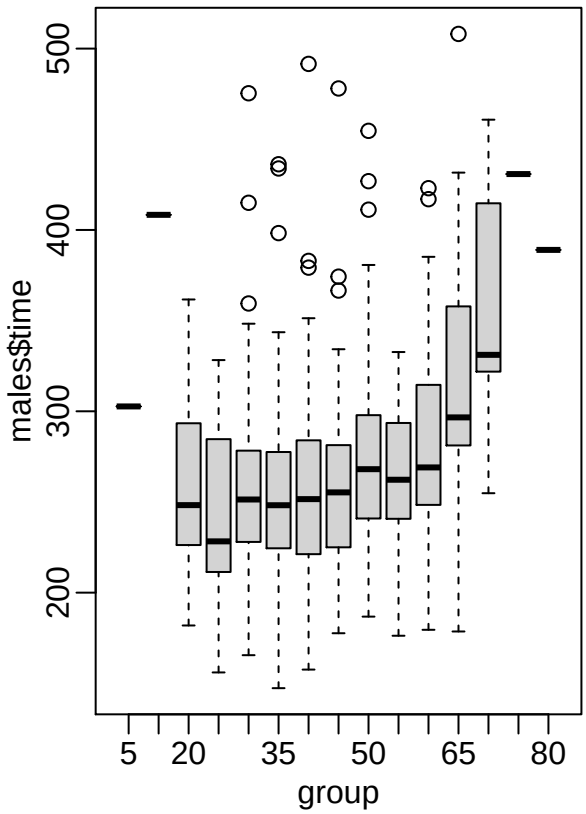
```
par(mfrow = c(1,2))
females <- filter(nym.2002, gender=="Female") %>% select(time,age)
males <- filter(nym.2002, gender=="Male") %>% select(time,age)
female_correlation <- cor(females$age,females$time, method="pearson")
male_correlation <- cor(males$age,males$time, method="pearson")

plot(females$age,females$time, xlab="Female age", ylab="Female time", main=paste0("Correlation=",female
plot(males$age,males$time, xlab="Male age", ylab="Male time", main=paste0("Correlation=",male_correlati
```



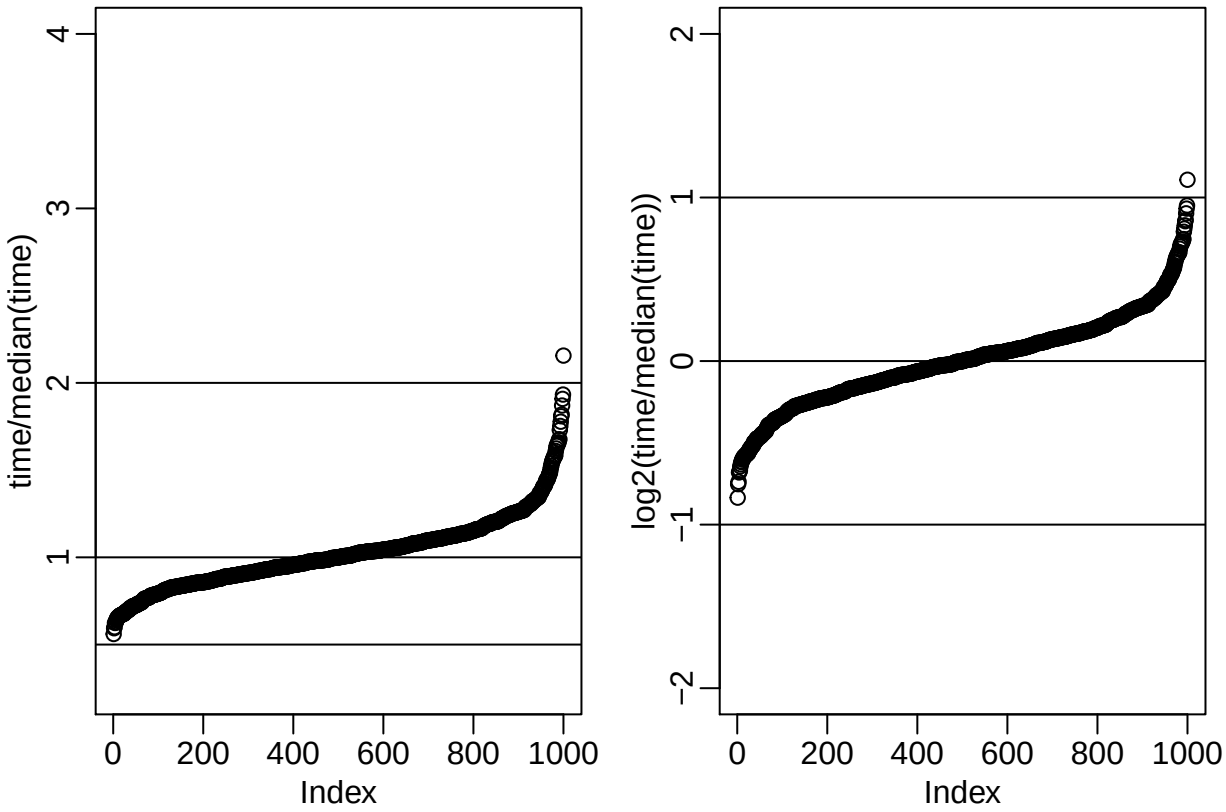
If we interpret these correlations and these scatter plots, we would conclude that the older we get, the slower we run marathons, regardless of gender.

```
library(rafalib)
mypar(1,2)
group <- floor(females$age/5) * 5
boxplot(females$time~group)
group <- floor(males$age/5) * 5
boxplot(males$time~group)
```






```
time <- sort(nym.2002$time)
mediantime <- time[round(length(time)/2)]
bestime <- time[length(time)]


mypar(1,2)
plot(time/median(time), ylim=c(1/4,4))
abline(h=c(1/2,1,2))
plot(log2(time/median(time)),ylim=c(-2,2))
abline(h=-1:1)
```



If we use box plots to visualize times with respect to age, we can see that finish times are constant up through around 50-60, then they get slower.