# Descriptive Statistics: the Grandmother's tale edition

M P Gururajan, Hina A Gokhale and Dayadeep Monder

Indian Institute of Technology Bombay, Mumbai

In this session, we are going to use a very simple data set for our analysis. The dataset is called `RKNGT.csv` and is stored in `csv` format. The name RKNGT stands for *R K Narayan* and his book Grandmother's tale. The data lists the number of words in the first hundred sentences of the book.

Of course, we know how to load the data set. Before we do that, we will also remove all data from the previous sessions. And, after reading the data, we will check it once before proceeding further.

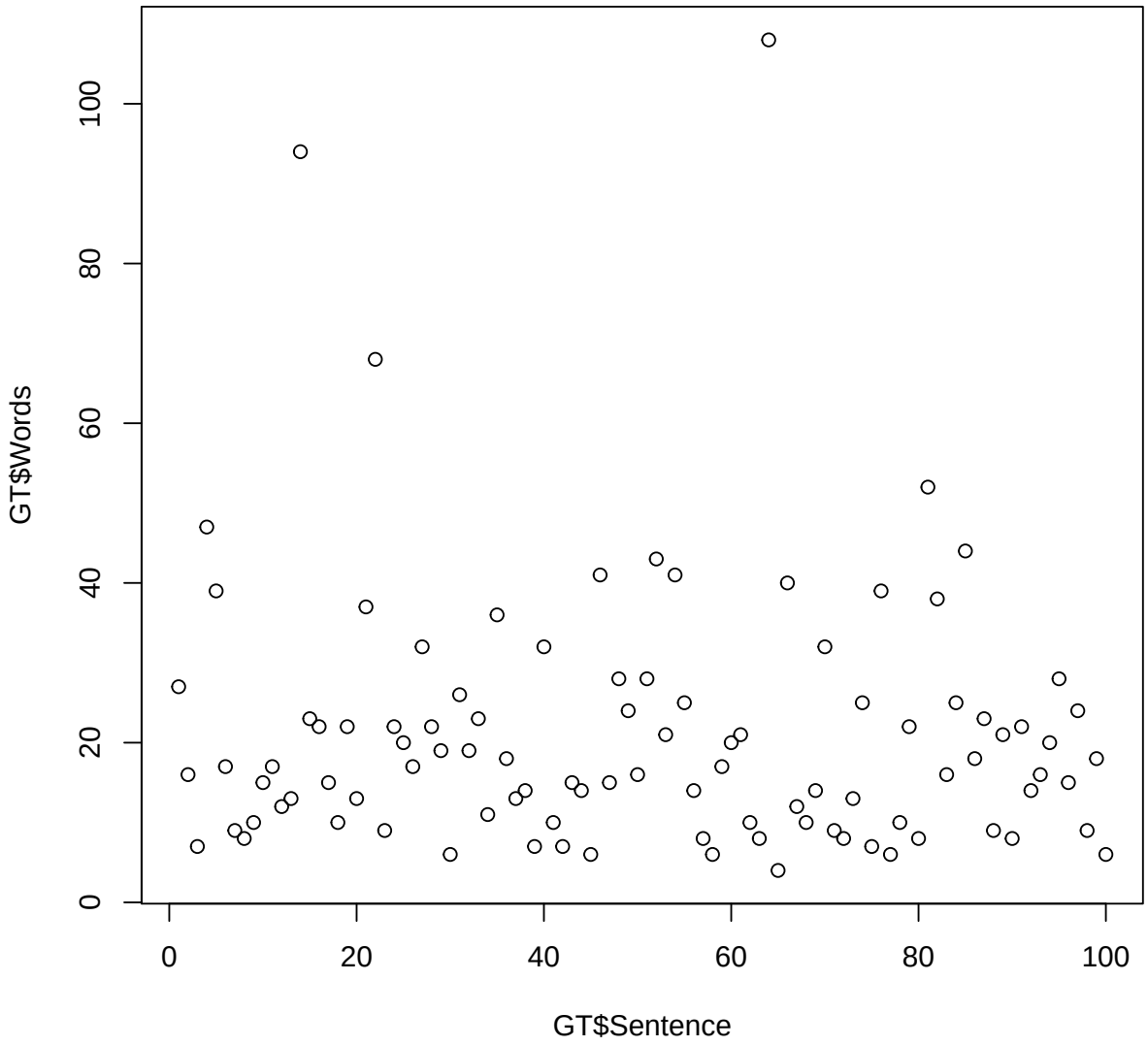
```
unlink("~/.RData")
GT <- read.csv("../Data/RKNGT.csv",header=TRUE)
str(GT)

## 'data.frame': 100 obs. of  2 variables:
##  $ Sentence: int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Words   : int  27 16 7 47 39 17 9 8 10 15 ...
```

## 1 The graphical measures!

As indicated in my lecture, let us first explore the data set by plotting it in various ways before calculating any numerical measures of the data!

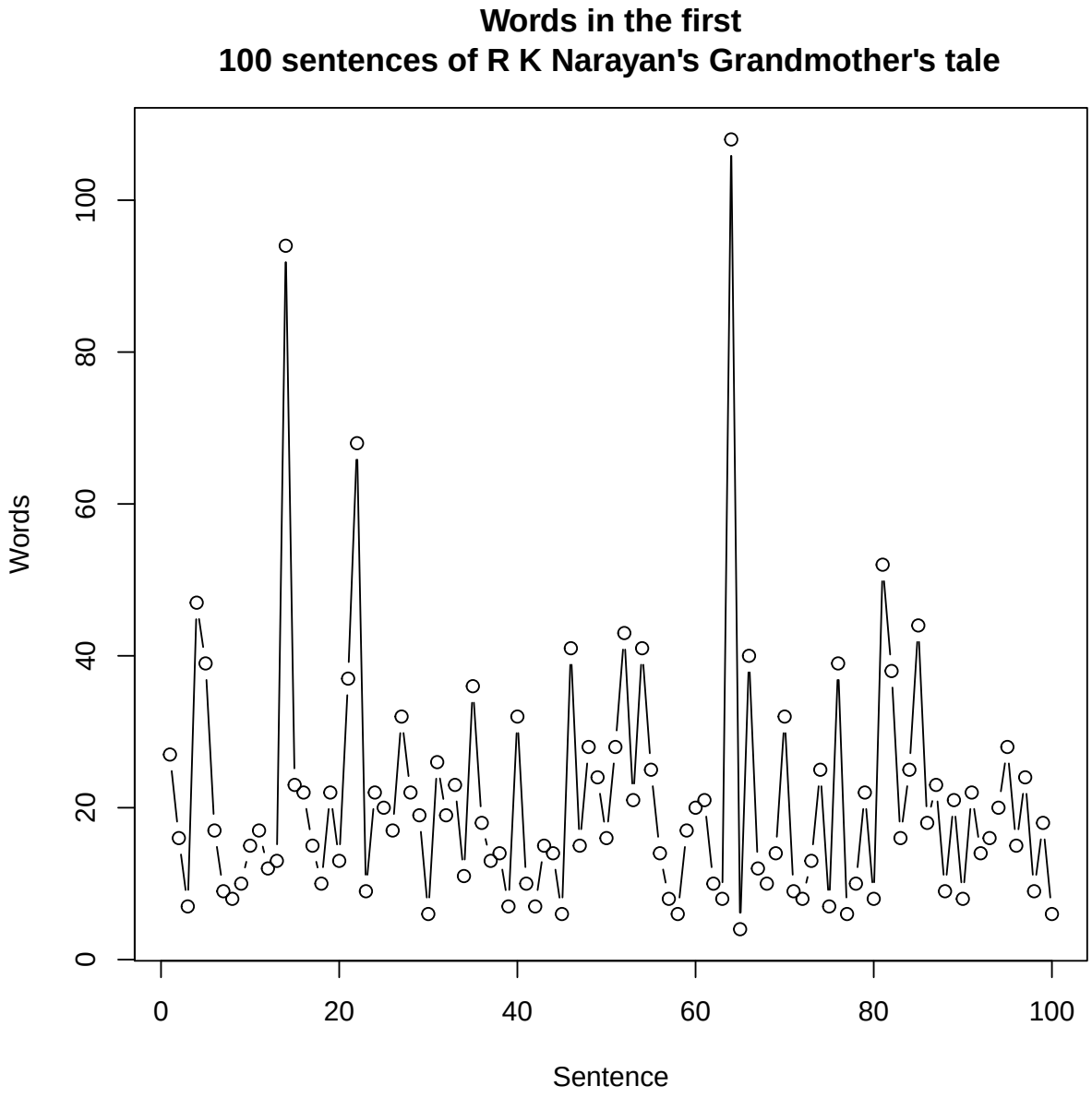
The first and easiest thing to do is to plot the data:

```
plot(GT$Sentence,GT$Words)
```




We can clean it up a bit by making the labels of the axes proper and giving a title to the plot. In addition, we will also add a line to the plot points to show the mix of short and long sentences in R K Narayan's writing:

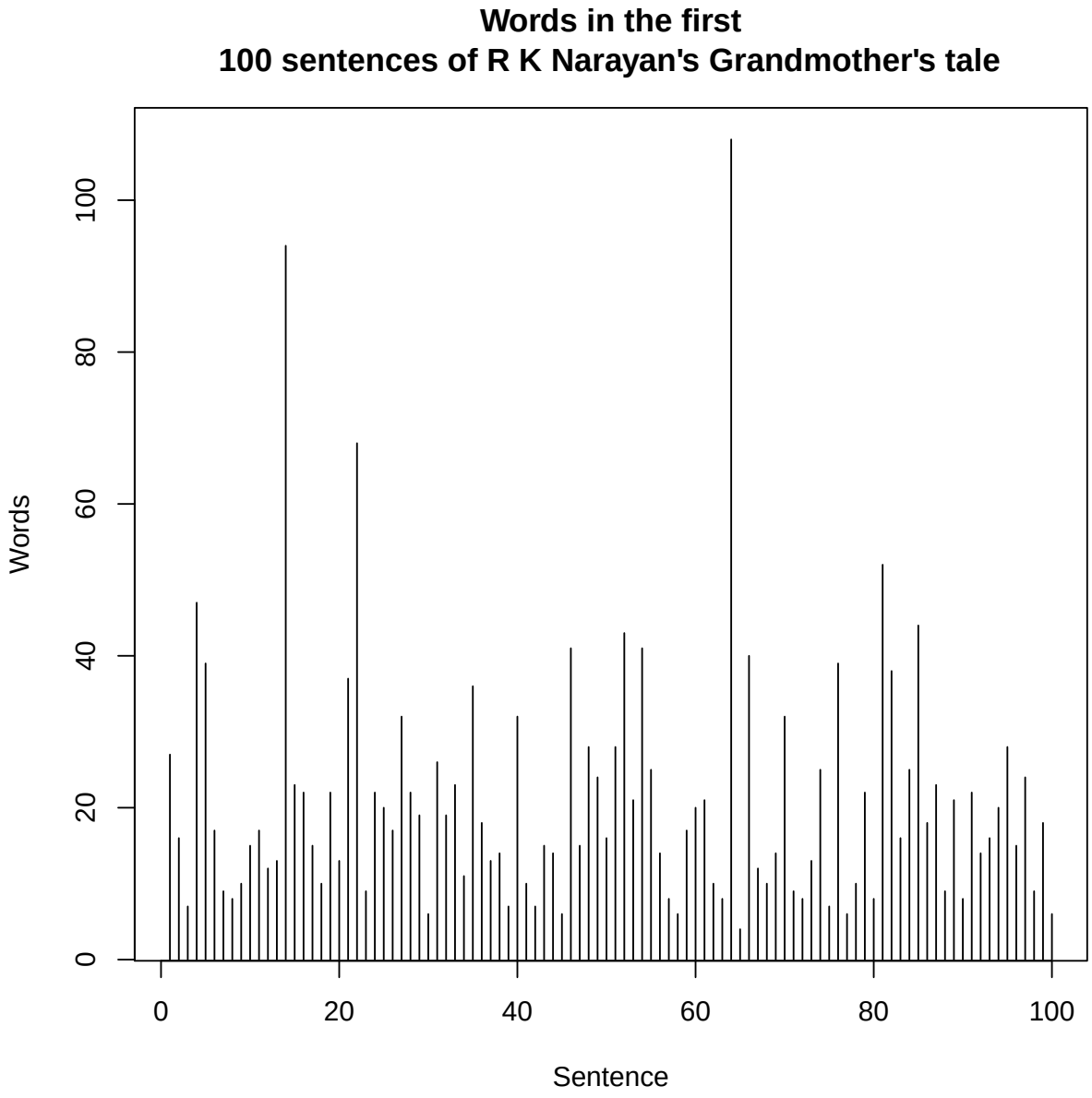
```
plot(GT$Sentence,GT$Words, type="b",
xlab="Sentence",ylab="Words",main="Words in the first
100 sentences of R K Narayan's Grandmother's tale")
```




Can you guess the mean number of words per sentence from this plot?

Let us try one more type of plotting – this time around with bars!

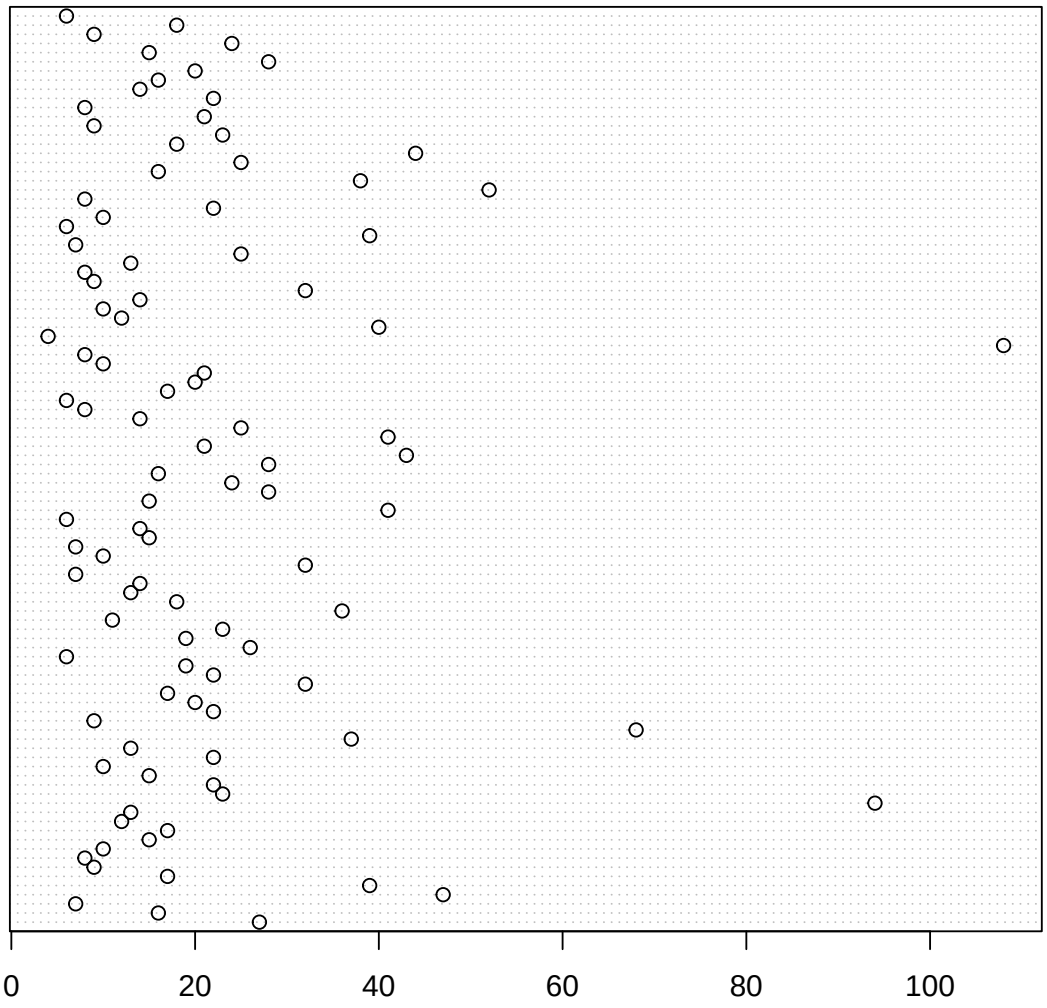
```
plot(GT$Sentence,GT$Words, type="h",
xlab="Sentence",ylab="Words",main="Words in the first
100 sentences of R K Narayan's Grandmother's tale")
```




Like the indicator on a music system, this plot brings out the rhythms of R K Narayan's writing clearly and evocatively!

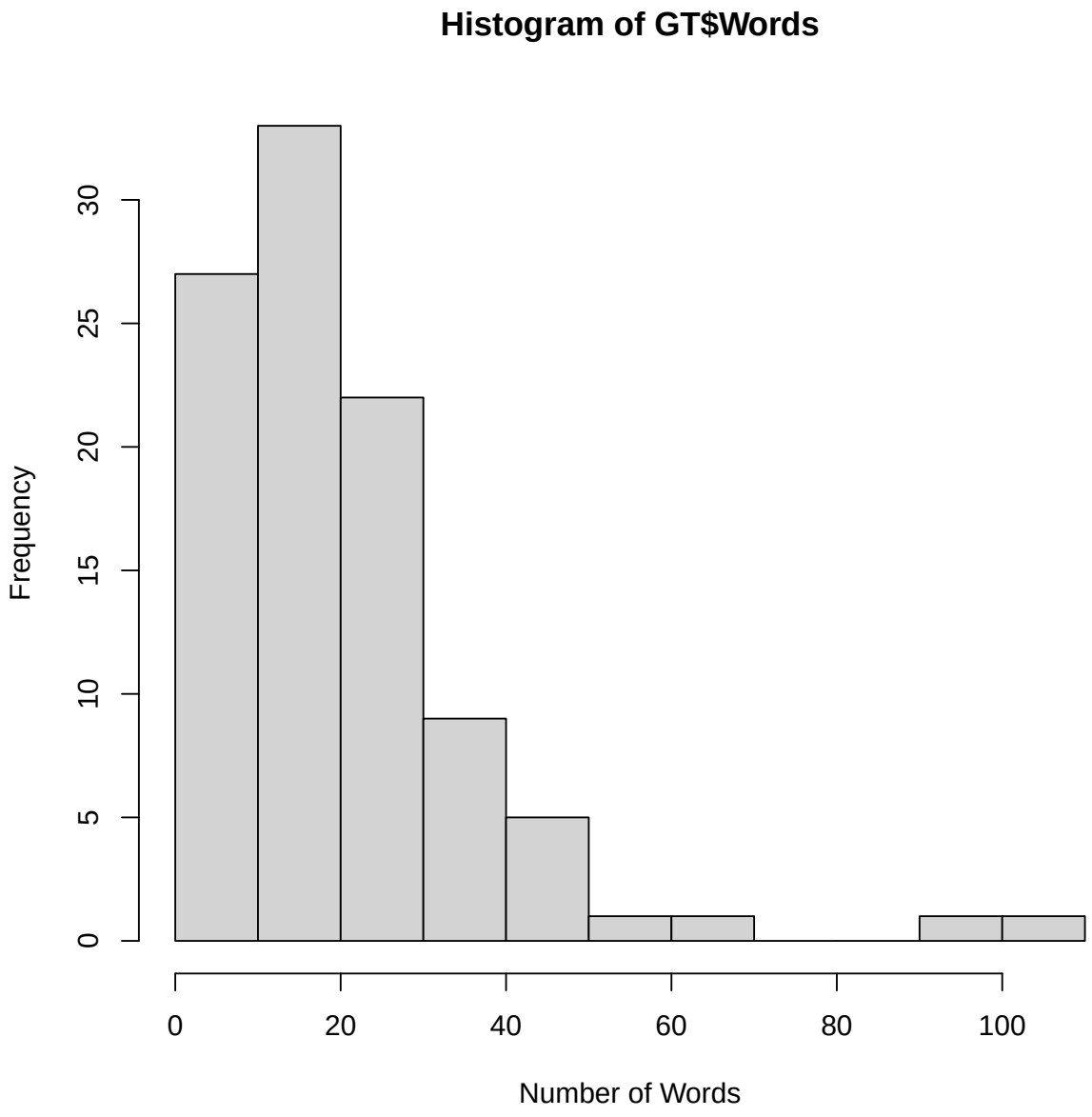
We can make a dotchart which indicates the frequency of occurrence – but at the cost of information on sequencing:

```
dotchart(GT$Words)
```



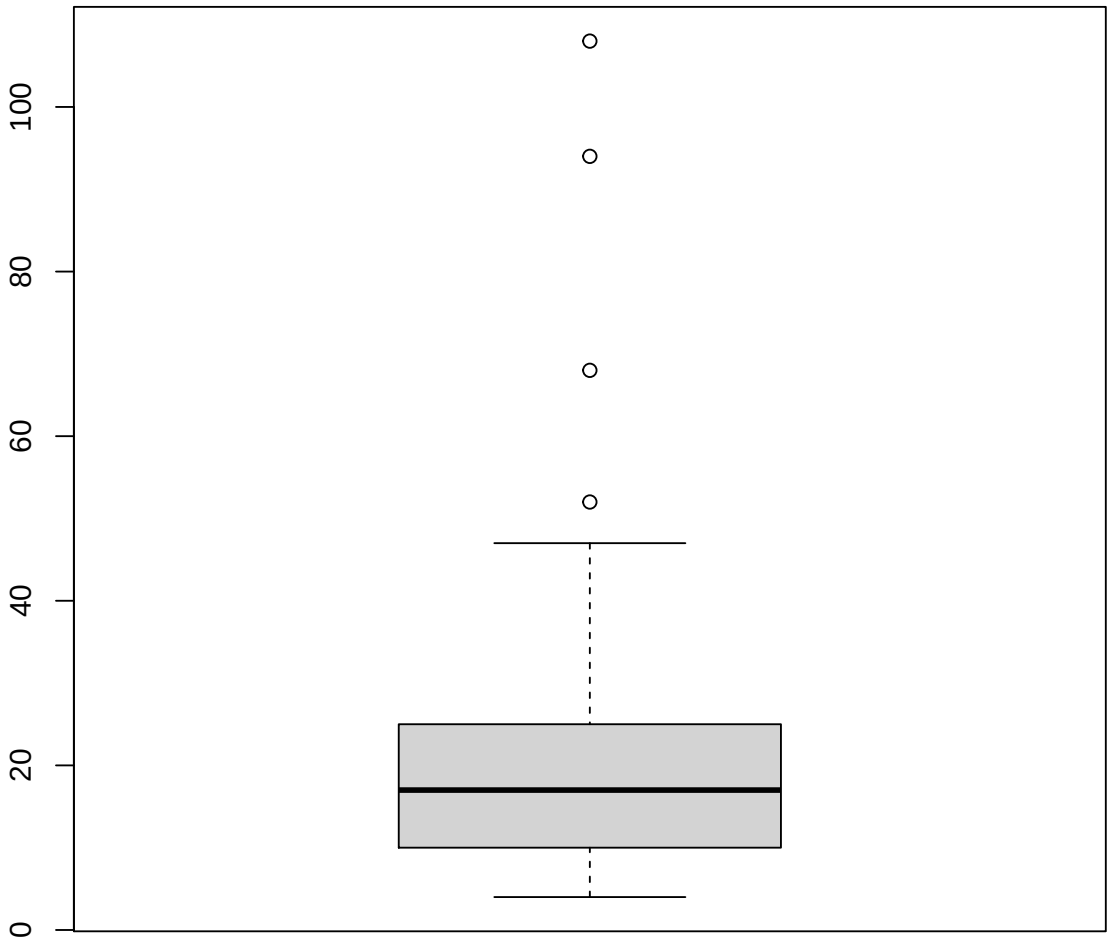

One can also draw a histogram of the data which gives the frequency of sentences with words in different ranges; note that here again the sequence is lost (and hence the information of how R K Narayan has mixed sentences of different length in his book). Of course, it is easy to see that the mean number of words per sentence is about 20. The distribution is skewed to the right. This would also imply that the median will be lower than the mean – we can verify these by calculating these quantities in the next section!

```
hist(GT$Words,xlab="Number of Words")
```



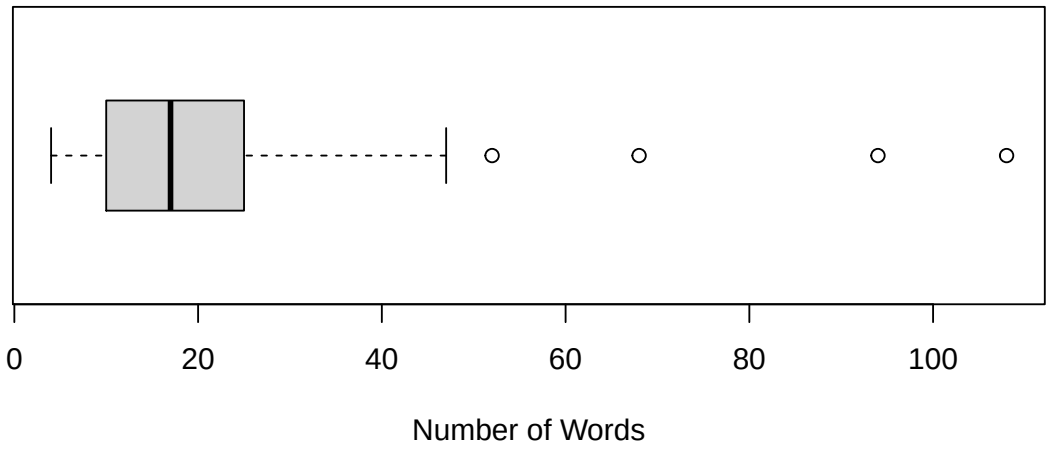

Another plot that includes some of the numerical descriptors is the boxplot – which, in addition to the data, also indicates through the whiskers (lines) and the boxes, the median and the range of the data – here, for example, one can clearly identify the outliers – unusually long sentences!

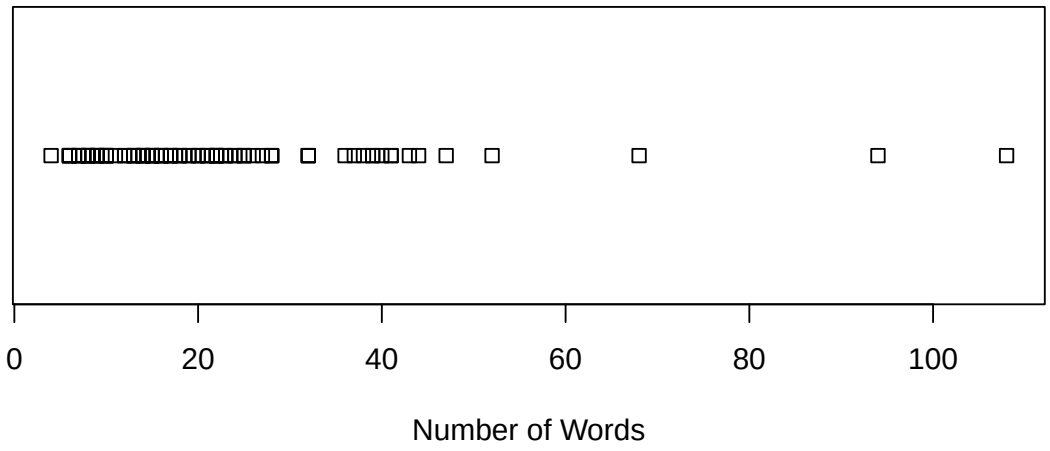
```
boxplot(GT$Words,xlab="Number of Words")
```




The boxplot can be plotted horizontally and let us compare it with the stripchart. To do this, I am going to plot two plots in a single figure.

```
par(mfrow=c(2,1))
boxplot(GT$Words,xlab="Number of Words",
        horizontal=TRUE)
stripchart(GT$Words,xlab="Number of Words")
```




## Home work

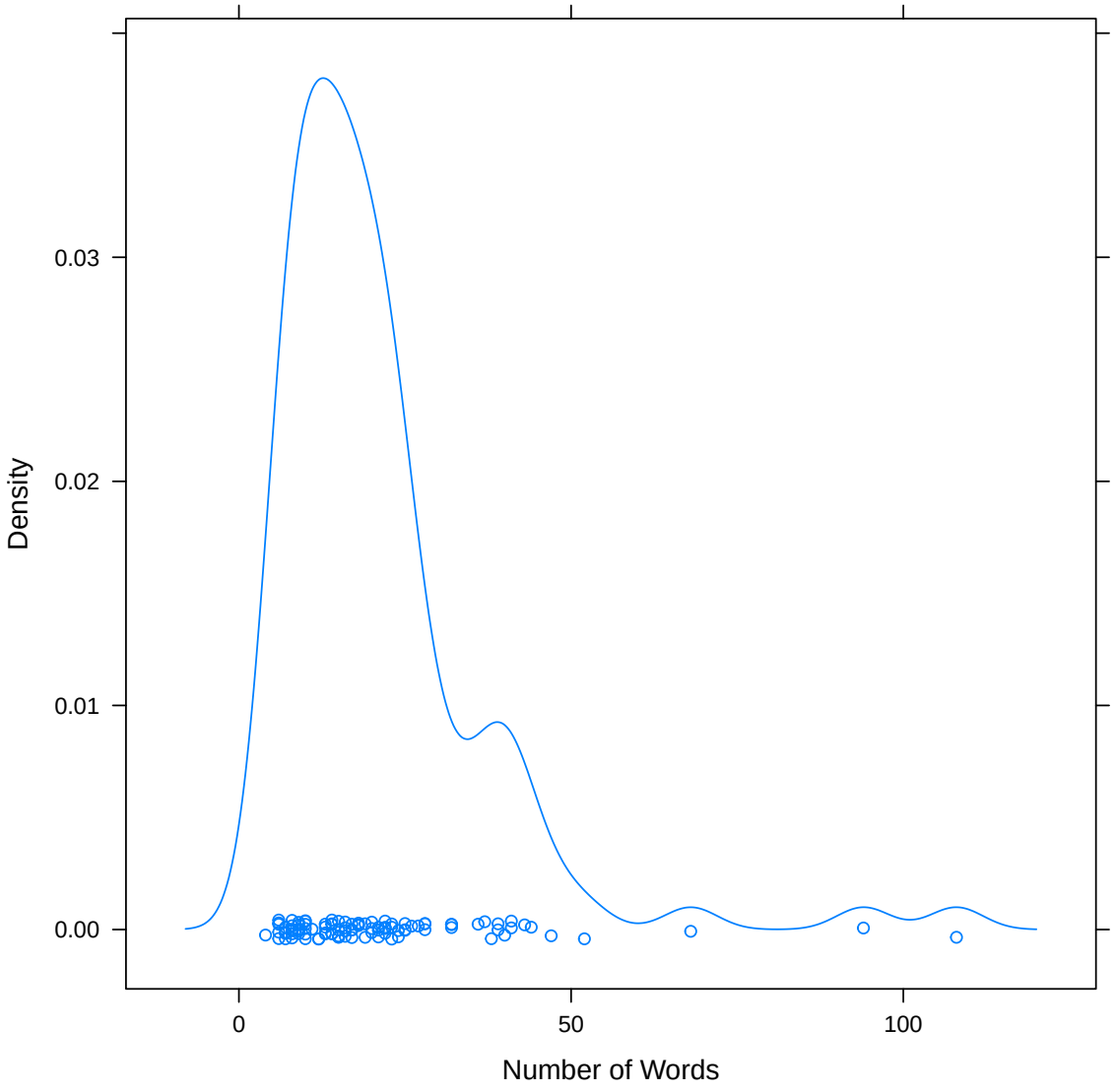
This probably is a good point for some home work!

- Check if the library `lattice` is already loaded in your version of R. You can check this using `library()` command.
- If it is not loaded, load the library. You can do this using `library(lattice)` command.
- Check if the package lattice is loaded. You can do this by using `search()` command.
- Now, use the command `densityplot` from `lattice` on the data and see what you get! Note that you have to make sure that you remove the two rows of plots that you opted for in the previous command (using `par(mfrow=c(1,1))`) before you do the new plot.

We will not interpret the density plot now! We will learn about frequency plots, cumulative frequency plots and density plots later!

**Answer to the last part of the home work**

```
## Warning in library():  libraries '/usr/local/lib/R/site-library', '/usr/lib/R/site-library'
contain no packages

##  [1] ".GlobalEnv"        "package:lattice"   "package:knitr"     "package:stats"
##  [5] "package:graphics"  "package:grDevices" "package:utils"     "package:datasets"
##  [9] "package:methods"   "Autoloads"         "package:base"
```




There are more plots such as stem-and-leaf plots and pie-charts, for example. We will learn them with the next tutorial. In the meanwhile, let us calculate some of the numerical descriptors of the data.

# 2 Numerical description

Let us calculate the mean, median and standard deviation:

```
mean(GT$Words)

## [1] 21.18

median(GT$Words)

## [1] 17

sd(GT$Words)

## [1] 16.45034
```

Now, if I ask you "What is the average number of words per sentence in R K Narayan's writing?", it is wrong to say 21.18. Why? Because there is no sentence with .18 words. So, the correct answer is either 21 (rounded), or 22 (rounded conservatively to the higher value).

Given that the standard deviation is 16.45 (again, conservatively, 17), the complete and correct answer to the question is $22 \pm 17$. In other words, R K Narayan's sentences vary in size; may be it adds to the charm of his writing!

In some contexts, the same information is given in another fashion. For example, you can say that on the average, 1000 sentences of R K Narayan will contain about $2118 \pm 1645$ words. This is just another way of avoiding the fractional words. But, unless we are sure that the numbers beyond the decimal point have meaning, we should not use them.

In most of science and engineering, if the original data is whole number and did not carry any digits following the decimal point, it probably makes little sense to report the average and standard deviation to several decimal points; it is more meaningful to report them also without any digits after the decimal point.

Of course, the mean is about 20 as we guessed and the median is smaller than the mean as would be expected given the skew of the data to the right!

There are other quantities that one can calculate, namely,

- Range (`range`)
- Minimum and maximum values (`min`, `max`)
- Variance (`var`), and,
- Quantiles (`quantile` or `fivenum`)

Quantiles indicate the percentage of data points that lie below the given value; thus, in this data, for example, there are no sentences with word length below 4; there are about 25% of the sentences with word length below 10; and so on; there are no sentences with word length about 108. Of course one can also use the command `summary` which gives a combination of central and range related quantities.

## 2.1 Self-evaluation question

What does the command `qauntile(GT$Words, 0.9)` give?

Now, we are ready for the next edition – namely, the Marital status edition of the descriptive statistics. We will do all the analysis listed above (as a tutorial / home exercise) followed by some such as pie-chart which we have not yet done as a tutorial!!