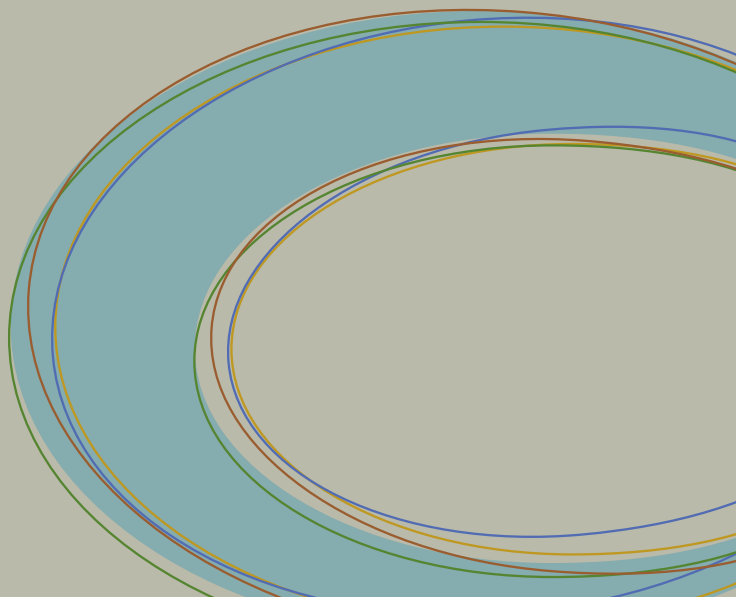
ΠΡΟΛΟΓΟΣ 2ΗΣ ΕΚΔΟΣΗΣ



Όπως προκύπτει από τον τίτλο του, βασικός σκοπός του συγκεκριμένου βιβλίου είναι να αποτελέσει ένα εγχειρίδιο αναφοράς για τη χρήση της γλώσσας προγραμματισμού R στην επεξεργασία και την ανάλυση δεδομένων. Από αυτήν την άποψη, η πρόκληση που αντιμετωπίζει είναι να κρατήσει σωστή ισορροπία μεταξύ τεχνικής και μεθοδολογίας, περιγραφής της γλώσσας και εφαρμογής της σε πραγματικά προβλήματα.

Η εμπειρία μου από τη διδασκαλία των αντικειμένων αυτών τα τελευταία δώδεκα χρόνια, μου έχει διδάξει ότι, όπως συμβαίνει και με τις φυσικές γλώσσες, έτσι και ο προγραμματισμός είναι προτιμότερο να διδάσκεται μέσα από συνεχή παραδείγματα και εφαρμογές. Τη λογική αυτή έχω προσπαθήσει να εφαρμόσω σε όλη την έκταση αυτού του βιβλίου. Η περιγραφή της R ξεκινάει από τις πιο βασικές αρχές και βήμα βήμα ο αναγνώστης μπορεί να χτίσει ένα σύνολο δεξιοτήτων που θα του επιτρέψουν να περάσει στην εφαρμογή της σε πιο πολύπλοκα εννοιολογικά πεδία όπως η στατιστική επαγωγή και η μοντελοποίηση δεδομένων.

Η δεύτερη αυτή έκδοση του βιβλίου «Ανάλυση Δεδομένων με την R» έρχεται, τέσσερα χρόνια μετά την πρώτη, να συμπληρώσει και να επεκτείνει το υλικό, καλύπτοντας μια ευρύτερη γκάμα τόσο τεχνικών όσο και θεωρητικών θεμάτων. Όλα τα κεφάλαια που περιέχονταν στην πρώτη έκδοση έχουν επικαιροποιηθεί, ενώ έχουν προστεθεί και πέντε νέα. Έτσι, το βιβλίο που κρατάτε στα χέρια σας περιέχει μια αναλυτική εισαγωγή στη σουίτα βιβλιοθηκών του tidyverse, που κερδίζει όλο και περισσότερο έδαφος μεταξύ των χρηστών, σε σχέση με την βασική έκδοση της R. Με την ίδια λογική, ένα νέο κεφάλαιο περιγράφει τη σύνταξη της βιβλιοθήκης γραφικών ggplot2, η οποία χρησιμοποιείται στη συνέχεια του βιβλίου σε εναλλαγή και παράλληλα με τις γραφικές παραστάσεις της βασικής R. Δύο νέα κεφάλαια παρουσιάζουν αρχές σύνταξης προγραμμάτων με την R για αριθμητικά αλλά και αλφαριθμητικά δεδομένα (strings). Ένα νέο κεφαλαίο, τέλος, περιγράφει τις βασικές αρχές της θεωρίας δικτύων και την ανάλυσή τους με τη βιβλιοθήκη igraph.

Σκοπός αυτής της δεύτερης έκδοσης ήταν εξαρχής να διατηρήσει ανοιχτή επικοινωνία με τους αναγνώστες του βιβλίου. Προσπάθησα να εντάξω σε αυτήν νέα στοιχεία και πιο διευρυμένη θεματολογία. Με αυτήν την λογική, οι αναγνώστες καλούνται να εκτιμήσουν κατά πόσο αυτός ο σκοπός επετεύχθη αλλά και να συνεχίσουν να αποτελούν τους εγκυρότερους κριτικούς του βιβλίου, με συμβουλές και επισημάνσεις για τυχόν σφάλματα και παραλείψεις.

ΚΕΦΑΛΑΙΟ

6

ΓΡΑΦΙΚΕΣ ΠΑΡΑΣΤΑΣΕΙΣ

Εισαγωγή

Μέχρι αυτό το σημείο, έχουμε παρουσιάσει και συζητήσει τρόπους για τον χειρισμό δεδομένων, τη δημιουργία συνόψεων, τη λήψη υποσυνόλων και την αναδιαμόρφωση δομών δεδομένων. Ωστόσο, όπως στο σινεμά, έτσι και στην ανάλυση δεδομένων μία από τις βασικές αρχές είναι η «μη το λες, δείξ' το» και έτσι, η ανάλυση των δεδομένων οφείλει, στη συνέχεια, να περάσει στη φάση της οπτικοποίησής τους και της γραφικής τους αναπαράστασης.

Ένα από τα βασικά πλεονεκτήματα της R σε σχέση με άλλες γλώσσες προγραμματισμού είναι οι σχεδόν απεριόριστες δυνατότητες που παρέχει στους χρήστες στη δημιουργία γραφικών. Οι περισσότεροι χρήστες αρχίζουν να χρησιμοποιούν την R με σκοπό να δημιουργήσουν πλούσια σε πληροφορία και υψηλής αισθητικής γραφήματα και, πράγματι, η R διαθέτει μια τεράστια εργαλειοθήκη λειτουργιών και πακέτων για την απεικόνιση δεδομένων μεγάλου όγκου και πολυπλοκότητας, που συμπεριλαμβάνουν χρονοσειρές, χάρτες, πολυδιάστατους πίνακες κ.ά. Το πιο σημαντικό είναι ότι αυτή η εργαλειοθήκη επεκτείνεται συνεχώς καθώς η R είναι γλώσσα ανοιχτού κώδικα και, έτσι, παρέχει μια ποικιλία διαρκώς ανανεούμενων, νέων λειτουργιών για γραφικά στα αποθετήριά της.

Ο τρόπος δημιουργίας γραφημάτων της R διαφέρει από τα προγράμματα που ενδεχομένως έχετε χρησιμοποιήσει έως τώρα (όπως το Excel ή το Graphpad) με την έννοια ότι δεν βασίζεται σε ένα γραφικό περιβάλλον αλλά σε γραμμή εντολών. Αυτό μπορεί να φαίνεται σαν μειονέκτημα στην αρχή, αλλά σύντομα θα δείτε ότι επιτρέπει πολύ εύκολα την ενσωμάτωση εντολών παραγωγής γραφικών σε προγράμματα και πιο πολύπλοκες υπολογιστικές διαδικασίες. Έτσι, μόλις κανείς εξοικειωθεί με τη διαδικασία κλήσης των συναρτήσεων γραφικών και την προσαρμογή τους στη γραμμή εντολών μπορεί να αποθηκεύσει τα πάντα σε προσαρμοσμένα σενάρια που θα παράγουν κομψά και πολύπλοκα γραφήματα, στην κυριολεξία με το πάτημα ενός κουμπιού. Ακόμη και τα πιο κομψά γραφήματα βασίζονται σε μερικούς πολύ απλούς κανόνες. Θα τα γνωρίσουμε ξεκινώντας από τις πιο βασικές γραφικές παραστάσεις και θα προσθέσουμε βαθμίδες «κομψότητας» βήμα-βήμα.

Ξεκινώντας από αυτό, και για τα δύο επόμενα κεφάλαια, θα δούμε πώς μπορούμε να χρησιμοποιήσουμε την R για τη γραφική αναπαράσταση δεδομένων. Αρχικά, με τη χρήση βασικών συναρτήσεων και στη συνέχεια με πιο εκλεπτυσμένες βιβλιοθήκες του *tidyverse* που στηρίζονται στη λεγόμενη γραμματική των γραφικών (grammar of graphics).

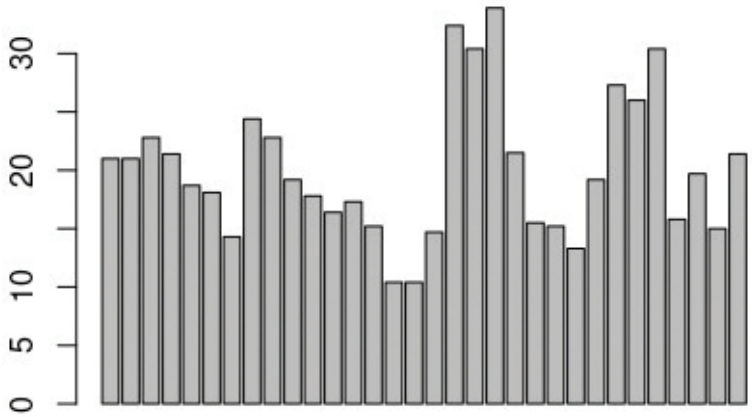
6.1 Ραβδογράμματα με την *barplot()*

Ας ξεκινήσουμε υποθέτοντας ότι έχετε έναν πίνακα αριθμητικών δεδομένων και θέλετε να δείτε πώς κάθε στοιχείο κυμαίνεται έναντι των υπολοίπων. Σε μια τέτοια περίπτωση, ένα απλό ραβδόγραμμα θα είναι αρκετό για να μας δώσει μια εικόνα. Η συνάρτηση ραβδογράμματος είναι η *barplot()*. Θα δημιουργήσουμε ένα απλό ραβδόγραμμα σε ένα μικρό σετ δεδομένων από το data frame *mtcars* που μπορούμε να φορτώσουμε από τη βασική έκδοση της R:

```
str(mtcars)
## 'data.frame':    32 obs. of  11 variables:
##  $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
##  $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
##  $ disp: num  160 160 108 258 360 ...
##  $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
##  $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
##  $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
##  $ qsec: num  16.5 17 18.6 19.4 17 ...
##  $ vs  : num  0 0 1 1 0 1 0 1 1 1 ...
##  $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
##  $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
##  $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```

Ένα ραβδόγραμμα χειρίζεται δεδομένα που βρίσκονται οργανωμένα σε ένα αριθμητικό διάνυσμα. Με την παρακάτω εντολή, θα δημιουργήσουμε ένα ραβδόγραμμα για ένα από τα διανύσματα του mtcars, το mpg:

```
barplot(mtcars$mpg)
```



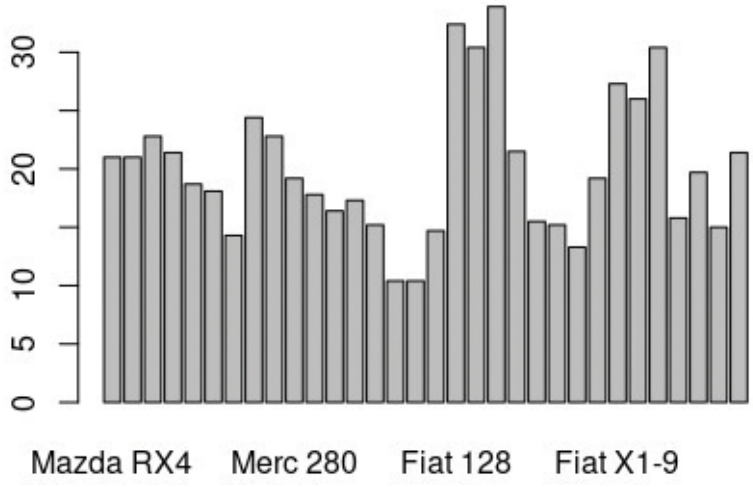
Εικόνα 6.1

Η κλήση της συνάρτησης δεν έχει καμία διαφορά από αυτήν όλων των συναρτήσεων που έχουμε δει έως τώρα με το αντικείμενο πάνω στο οποίο θα δράσει η συνάρτηση μέσα σε παρένθεση. Το αποτέλε-

σμα της κλήσης είναι μια γραφική παράσταση στην οποία κάθε στοιχείο του διανύσματος mpg αντιστοιχεί σε μία στήλη. Με την κλήση της συνάρτησης *barplot()* (ή οποιασδήποτε συνάρτησης σχεδίασης) ένα νέο παράθυρο θα ανοίξει στο περιβάλλον της R με το δημιουργημένο γράφημα (αν χρησιμοποιείτε το RStudio, μπορείτε να το δείτε σε ξεχωριστό παράθυρο). Το παραπάνω γράφημα φαίνεται αρκετά φτωχό τόσο αισθητικά όσο και πληροφοριακά. Εκτός από τον κάθετο άξονα, στον οποίο φαίνεται το ύψος των αριθμητικών τιμών, δεν υπάρχουν άλλα στοιχεία που θα μας βοηθήσουν να καταλάβουμε τι αναπαρίσταται. Χρειαζόμαστε περισσότερες πληροφορίες και επιλογές για τους άξονες, τίτλους, υπόμνημα κ.λπ. Όλες οι συναρτήσεις γραφικών στην R προβλέπουν έναν μεγάλο αριθμό παραμέτρων που μας επιτρέπουν να καθορίσουμε όλα τα παραπάνω. Ας δούμε πώς μπορούμε να προσθέσουμε επιμέρους στοιχεία στο γράφημα, βήμα-βήμα:

1. Τίτλο δεδομένων στον οριζόντιο άξονα (names=):

```
barplot(mtcars$mpg, names=rownames(mtcars))
```

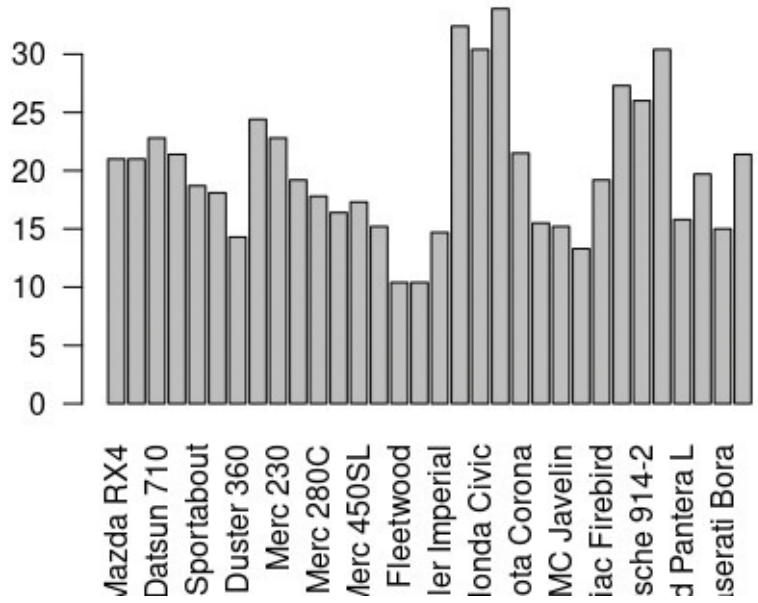


Εικόνα 6.2

Με την παραπάνω επιλογή, έχουμε ζητήσει από την *barplot()* να αποδώσει τα ονόματα των μοντέλων κάτω από τις μπάρες που αντιστοιχούν στις τιμές τους. Βλέπουμε όμως ότι το πλήθος τους είναι τέτοιο που δεν επιτρέπει σωστή απεικόνιση, καθώς οι ετικέτες των ονομάτων υπερκαλύπτονται και καθίστανται έτσι δυσανάγνωστα. Υπάρχουν διάφοροι τρόποι για να ξεπεράσουμε κάτι τέτοιο. Ας δούμε παρακάτω πώς μπορούμε να κάνουμε:

2. Αλλαγή διάταξης δεδομένων στους άξονες (las=):

```
barplot(mtcars$mpg, names=rownames(mtcars), las=2)
```

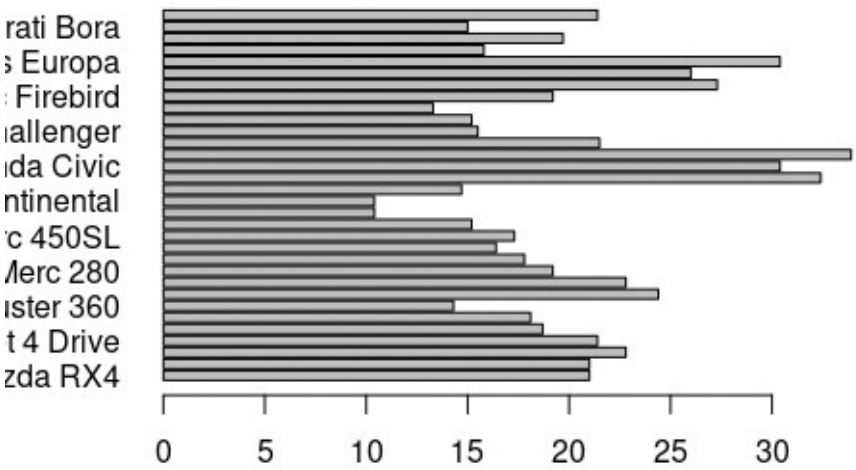


Εικόνα 6.3

Θέτοντας την παράμετρο *las* (labels) ίση με την τιμή 2, οι τιμές των δύο αξόνων γίνονται κάθετες προς αυτούς. Τώρα, τα ονόματα των μοντέλων φαίνονται κάπως καλύτερα, εξακολουθούν όμως να είναι αρκετά μεγάλα για την εικόνα και επιπλέον είναι δυσανάγνωστα λόγω του προσανατολισμού τους. Θα ήταν πολύ καλύτερα αν το ραβδόγραμμα σχηματιζόταν αντίστροφα με τις ράβδους να εκτείνονται οριζόντια:

3. Αλλαγή διάταξης ραβδογράμματος (horiz=):

```
barplot(mtcars$mpg, names=rownames(mtcars), las=1, horiz=T)
```



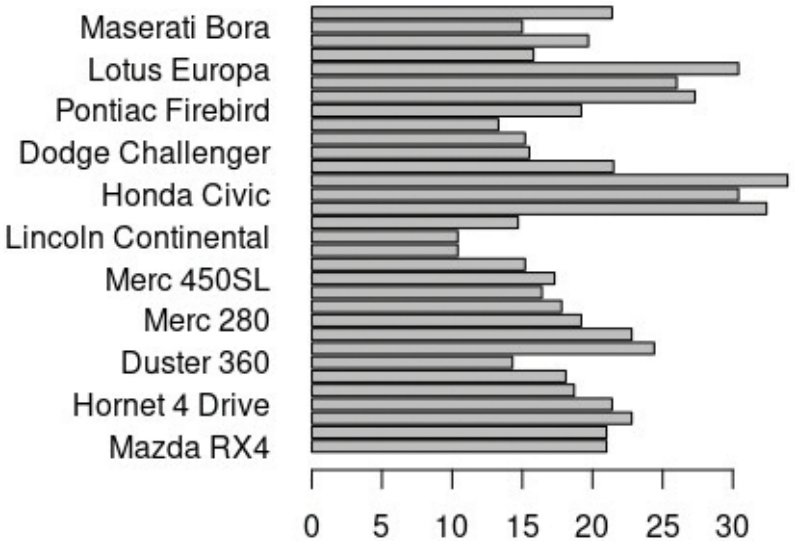
Εικόνα 6.4

Θέτοντας «αληθή» (T) τη λογική τιμή της παραμέτρου *horiz=LOGICAL*, επιτυγχάνουμε την αντιστροφή του προσανατολισμού, με τα δεδομένα τώρα να αναπαρίστανται με οριζόντιες μπάρες. Προσέξτε ότι έχουμε επαναφέρει το las στην τιμή 1 για να διαβάζουμε καλύτερα τον άξονα τον αριθμητικών τιμών (που τώρα είναι ο οριζόντιος). Το πρόβλημα εξακολουθεί να βρίσκεται στα περιθώρια (που είναι μικρά

και αποκρύπτουν τμήμα των ονομάτων) και στο μέγεθος της γραμματοσειράς που είναι πολύ μεγάλη. Ας δούμε πως μπορούμε να λύσουμε τα δύο αυτά προβλήματα, τροποποιώντας μερικές νέες παραμέτρους:

4. Αλλαγή περιθωρίων πλαισίου (par(mar)=c()):

```
par(mar=c(5,10,2,2))
barplot(mtcars$mpg, names=rownames(mtcars), las=1, horiz=T)
```



Εικόνα 6.5

Η αλλαγή που κάναμε επιτυγχάνεται με ξεχωριστή εντολή, περνώντας στη συνάρτηση *par()*, η οποία ορίζει παραμέτρους γραφικών, τον καθορισμό των περιθωρίων μέσω της *mar=c(bottom, left, top, right)*. Οι τέσσερις τιμές αντιστοιχούν στα σχετικά περιθώρια όπως φαίνονται παραπάνω. Για τις ανάγκες της δικής μας γραφικής, θέλουμε μια μεγάλη τιμή στο αριστερό περιθώριο, γι' αυτό το θέτουμε ίσο με 10. Αξίζει να επισημάνουμε εδώ ότι οι αλλαγές μέσω της *par()* είναι πάντοτε παροδικές, έχουν δηλαδή περιορισμένη εμβέλεια που διαρκεί μέχρι την αμέσως επόμενη εκτέλεση της συνάρτησης δημιουργίας γραφικών. Για τον λόγο αυτό, θα πρέπει πάντα να επαναλαμβάνονται πριν την εκτέλεση της γραφικής συνάρτησης. Ας προχωρήσουμε σε μερικές ακόμα τροποποιήσεις, μικραίνοντας τη γραμματοσειρά των μοντέλων αυτοκινήτων και δίνοντας τίτλους στο γράφημα και στους άξονες.

5. Τίτλοι αξόνων και γραφήματος (xlab/main/cex):

```
par(mar=c(5,10,2,2))
barplot(mtcars$mpg, names=rownames(mtcars), las=1, horiz=T,
cex.names=0.6, xlab="mpg", main="MtCars Miles per Galon")
```