

αριθ

Εργασία 1

Θέμα Εργασίας: Εισαγωγή στην R και Περιγραφική Στατιστική

Παναγιώτης Βλάχος

30/03/2018

Αρ. Μητρώου : ge15124

Σχολή : ΕΜΦΕ

Εξάμηνο : 6^ο

Άσκηση 1

Για την επεξεργασία και την μελέτη του προβλήματός μας, θα χρησιμοποιήσουμε το στατιστικό πακέτο R, μέσω του οποίου θα καταφέρουμε να συλλέξουμε την απαιτούμενη πληροφορία και να παρουσιάσουμε τα ζητούμενα στατιστικά αποτελέσματα. Για κάθε ενέργεια που εκτελούμε στην R(πληκτρολόγηση εντολών) θα υπάρχει σχετική αναφορά αλλά και κάποια σχόλια γι' αυτήν, προκειμένου να γίνεται αντιληπτός ο λόγος για τον οποίο χρησιμοποιήθηκε αλλά και για να γίνεται σαφή η επεξήγηση της λειτουργίας της.

Ερώτημα 1

Ξεκινώντας, επιθυμούμε να περάσουμε στην R τα δεδομένα τα οποία είναι αποθηκευμένα στο δοθέν αρχείο(υπό μορφή excel :Diamond.csv) .Για τον λόγο αυτό θα χρησιμοποιήσουμε την εντολή read.table(.) ως εξής:

```
> file.choose()
```

```
[1] "C:\\Users\\Chris\\Downloads\\Diamond.csv"
```

Όπου συνδέουμε το αρχείο με την R.

```
a<read.table("C:\\Users\\Chris\\Downloads\\Diamond.csv",header=TRUE,sep=",",str  
ip.white=TRUE,na.strings="*",colClasses=c(rep("character",6)))
```

Όπου η εντολή read.table(.) διαβάζει ένα αρχείο σε μορφή πίνακα και δημιουργεί ένα πλαίσιο δεδομένων από αυτό. Η παράμετρος header=TRUE είναι μια λογική τιμή που υποδεικνύει αν το αρχείο περιέχει τα ονόματα των μεταβλητών στην πρώτη του γραμμή. Η sep="," δείχνει τον τρόπο με τον οποίο διαχωρίζονται οι χαρακτήρες στο αρχείο. Η na.strings="*" μετατρέπει τις αγνοούμενες τιμές που συμβολίζονται με «*» σε «Na».Ο τύπος του αντικειμένου που δημιουργείται είναι data frame.

Επίσης εκτελούμε και τις παρακάτω εντολές οι οποίες εξασφαλίζουν ότι οι συγκεκριμένες τιμές είναι αριθμητικές:

```
> class(a$id)="numeric"
```

```
> class(a$carat)="numeric"
```

```
> class(a$price)="numeric"
```

Ερώτημα 2

Έχουμε στη διάθεσή μας τώρα 5 μεταβλητές προς μελέτη. Αναλυτικά είναι : 1)η μονάδα βάρους σε καράτια (**carat**), το χρώμα (**colour**, με κατηγορίες τις "D", "E", "G", "F", "H", "I"), η καθαρότητα (**clarity**, με κατηγορίες τις "IF", "VS1", "VS2", "VVS1", "VVS2"), η πιστοποίηση (**certification**, με κατηγορίες τις "GIA", "IGI", "HRD") και η τιμή (**price**) σε δολάρια. Μια γρήγορη, αλλά ικανοποιητική περιγραφή μας δίνει η εντολή: `summary(a,na.rm=TRUE)` η οποία εμφανίζει στατιστικά μεγέθη όπως το πλήθος των ποσοτήτων στις διάφορες κατηγορίες, το πλήθος των «Να» στοιχείων αλλά και για τις αριθμητικές μεταβλητές τη μέση τιμή, τη διασπορά, τη μέγιστη και την ελάχιστη τιμή. Τα αποτελέσματα της εκτέλεσής της φαίνονται παρακάτω:

```
> summary(a,na.rm=TRUE)
```

Carat	Colour	Clarity	Certification	Price
Min. : 0.1800	D : 16	IF : 44	GIA : 149	Min. : 638
1 st Qu. : 0.3575	E : 44	VS1 : 80	HRD : 79	1 st Qu. : 1628
Median : 0.6200	F : 82	VS2 : 53	IGI : 78	Median : 4221
Mean : 0.6324	G : 65	VVS1 : 52	NA's : 2	Mean : 5031
3 rd Qu. : 0.8525	H : 61	VVS2 : 78		3 rd Qu. : 7524
Max. : 1.1000	I : 40	NA's : 1		Max. : 16008
NA's : 4				NA's : 1

Σημειώνουμε ότι οι ποσοτική μεταβλητή 1st Qu. ονομάζεται ενδοτεταρτημοριακό εύρος και δηλώνει τη παρατήρηση εκείνη που είναι μεγαλύτερη ή ίση από το 25% ακριβώς των παρατηρήσεων και αντίστοιχα η 3rd Qu. την παρατήρηση εκείνη που είναι μεγαλύτερη ή ίση από το 75% ακριβώς των παρατηρήσεων.

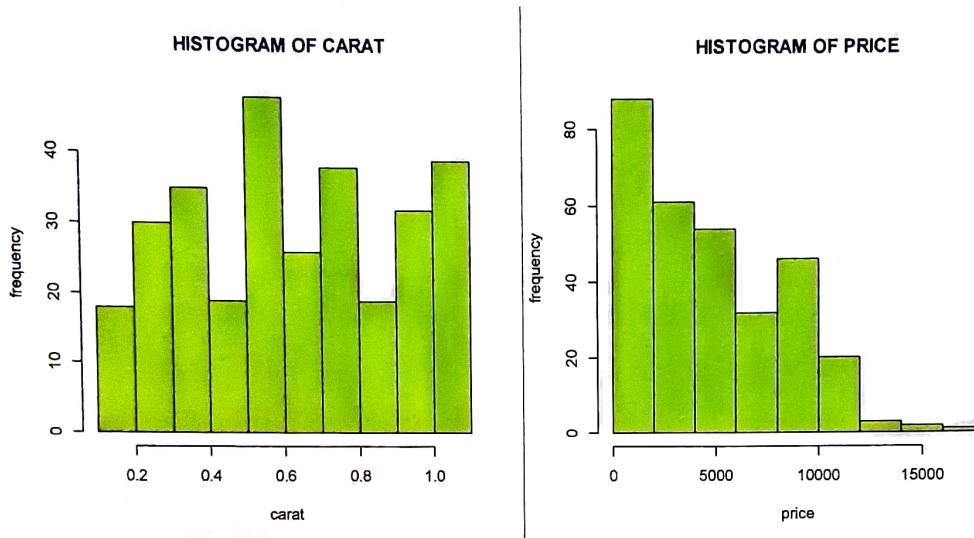
Παράλληλα, το μεγάλο μέγεθος του δείγματος που διαθέτουμε αλλά και ο μικρός σχετικά αριθμός των NA's στοιχείων δεν καταστεί αναγκαία την αντικατάστασή τους με κάποια δική μας εκτίμηση.

Συνεχίζοντας, φαίνεται χρήσιμη αλλά και αναγκαία η οπτικοποίηση των μεταβλητών μας, καθώς θα μας προσφέρει μια καλύτερη αίσθηση αυτών. Για τον λόγο αυτό θα χρησιμοποιήσουμε τις γραφικές δυνατότητες που μας προσφέρει η R. Ξεκινώντας την γραφική μελέτη μας με τις αριθμητικές μεταβλητές, θα δημιουργήσουμε δύο ιστογράμματα με την χρήση της εντολής `hist(x)`. Πιο συγκεκριμένα, κατασκευάζουμε τα ιστογράμματα των μεταβλητών "Carat" και "Price" ως εξής:

```
> hist(a$carat,xlab="carat",ylab="frequency",main="HISTOGRAM OF CARAT",col="darkolivegreen1")
```

```
> hist(a$price,xlab="price",ylab="frequency",main="HISTOGRAM OF PRICE",col="darkolivegreen1")
```

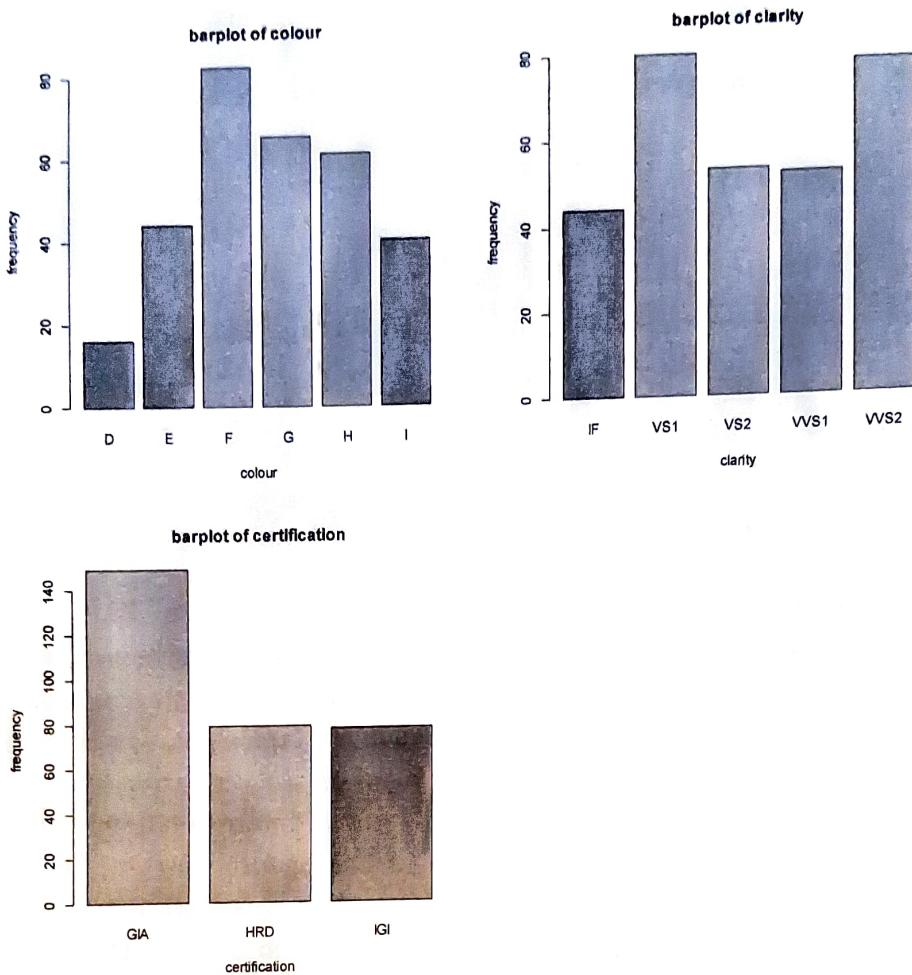
Και τα εμφανίζουμε παρακάτω με το αριστερό να είναι το ιστόγραμμα της Carat και αντίστοιχα το δεξί το ιστόγραμμα της Price.



Συνεχίζουμε τώρα και περνάμε στις κατηγορικές μεταβλητές οι οποίες είναι το χρώμα(colour), η καθαρότητα(clarity) και η πιστοποίηση(certification). Η επιλογή της αναπαράστασης είναι το ραβδόγραμμα (και όχι τομεόγραμμα) στη συγκεκριμένη περίπτωση καθώς μας δίνει καλύτερη πληροφόρηση και αυτό υλοποιείται με την barplot(table(A)). Στην δική μας περίπτωση γράφουμε:

```
> barplot(table(a$colour),xlab="colour",ylab="frequency",main="barplot of colour")
> barplot(table(a$clarity),xlab="clarity",ylab="frequency",main="barplot of clarity")
> barplot(table(a$certification),xlab="certification",ylab="frequency",main="barplot of certification")
```

Και τα αποτελέσματα φαίνονται παρακάτω:



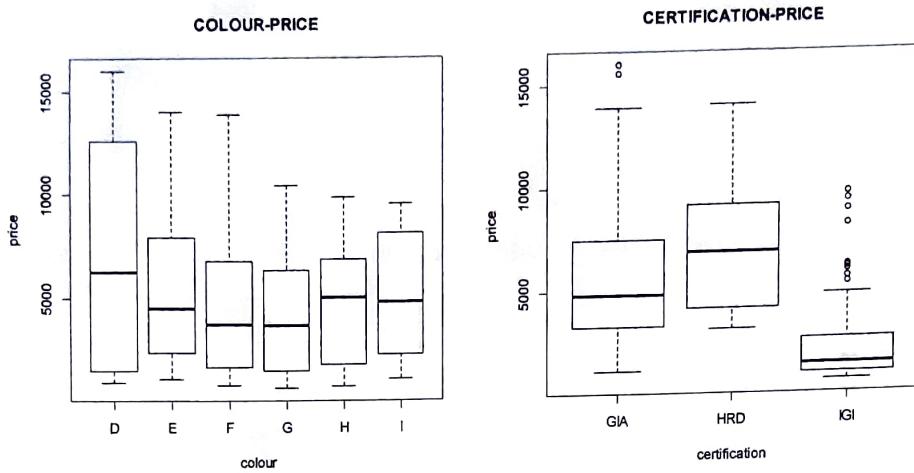
Ερώτημα 3

Αφού έχουμε ολοκληρώσει την μελέτη μας για τα γενικά χαρακτηριστικά του δείγματος, θα προχωρήσουμε στην διεξαγωγή πιο συγκεκριμένων συμπερασμάτων.

Θέλουμε να ελέγξουμε τη σχέση και την εξάρτηση μεταξύ της μεταβλητής "Price" τόσο με την Colour όσο και με την Certification. Με τη μορφή όμως των δεδομένων που διαθέτουμε δεν μπορούμε να βγάλουμε κάποιο άμεσο συμπέρασμα. Η γραφική αναπαράσταση θα μας βοηθήσει να επιτύχουμε ικανοποιητικά τον έλεγχο που επιθυμούμε. Έτσι με την χρήση θηλογράμματος (κατάλληλη επιλογή για σύγκριση) θα προσπαθήσουμε να λάβουμε την επιθυμητή πληροφορία. Οι εντολές που χρησιμοποιούμε για την δημιουργία αυτού είναι:

```
> plot(a$colour,a$price,xlab="colour",ylab="price",main="COLOUR-PRICE")
> plot(a$certification,a$price,xlab="certification",ylab="price"
,main="CERTIFICATION-PRICE")
```

Και παίρνουμε τα εξής:



Για το αριστερό θηκόγραμμα, παρατηρούμε πως υπάρχει μια μερική εξάρτηση της τιμής από το χρώμα. Αρχικά, βλέπουμε πως η μέση τιμή για οποιοδήποτε χρώμα δεν αλλάζει σημαντικά, καθώς κυμαίνεται γύρω από την τιμή 5000. Επίσης, ένα άλλο στοιχείο που μας οδηγεί στο συμπέρασμα της μη εξάρτησης αποτελεί το γεγονός ότι η ελάχιστες τιμές σχεδόν συμπίπτουν για κάθε χρώμα. Αντίθετα όμως από την παραπάνω διαπίστωση, οι μέγιστες τιμές είναι διαφορετικές καθώς το χρώμα "D", "E", "F" έχουν κατά πολύ μεγαλύτερες τιμές από τα υπόλοιπα χρώματα. Επίσης το "D" για ένα ικανοποιητικό μέγεθος τιμών έχει αρκετά υψηλή τιμή, με αποτέλεσμα να θεωρούμε ότι είναι μια πιο 'ακριβή' κατηγορία με βάση το χρώμα, χωρίς όμως να μπορούμε να πούμε με βεβαιότητα την πλήρη εξάρτηση των δυο μεταβλητών.

Για το δεξί θηκόγραμμα, μπορούμε με αρκετή σιγουριά να πούμε πως υπάρχει η εξάρτηση ανάμεσα στις δυο μεταβλητές. Αναλυτικότερα, η κατηγορία με την πιστοποίηση "GIA" όπως και η "HRD" έχουν μια ευελιξία στην τιμή τους, με ίσως την δεύτερη να είναι περισσότερο ακριβή ενώ αντίθετα η "IGI" έχει πολύ χαμηλές τιμές(εκτός από κάποιες μεμονωμένες περιπτώσεις).

Ερώτημα 4

Θέλουμε να δημιουργήσουμε μια νέα μεταβλητή η οποία θα έχει την δυνατότητα να μας δώσει πληροφορία για την μονάδα βάρους των διαμαντιών σε καράτια. Για τον σκοπό αυτό, θα κατασκευάσουμε την καινούρια μεταβλητή `best_diamond`, η οποία θα ξεχωρίζει τα διαμάντια σε αυτά με μονάδα βάρους σε καράτια μικρότερη του 0.51 και θα παίρνει την τιμή «`light`» γι' αυτά, ενώ για τα διαμάντια με μονάδα βάρους μεγαλύτερη του 0.51 θα παίρνει την τιμή «`heavy`». Ένας τρόπος για να κατασκευάσουμε στο λογισμικό μας τη μεταβλητή και να εκχωρήσουμε τις τιμές είναι η παρακάτω αλληλουχία εντολών:

```

> best_diamond=ifelse(a[["carat"]]>0.51,"heavy","light")

> best_diamond=factor(a$best_diamond)

> a$best_diamond=best_diamond

```

Με τον τρόπο αυτό δημιουργούμε μια νέα στήλη στον πίνακα a.

Αρχικά, δυο χρήσιμες πληροφορίες για την best_diamond είναι η συχνότητες και οι σχετικές συχνότητες εμφάνισης των τιμών «light» και «heavy» τις οποίες συλλέγουμε από τις εντολές:

```

> table(a$best_diamond)

> prop.table(table(a$best_diamond))

```

Και δημιουργούμε τον ζητούμενο πίνακα:

	ΣΥΧΝΟΤΗΤΕΣ	ΣΧΕΤΙΚΕΣ ΣΥΧΝΟΤΗΤΕΣ
“heavy”	196	0.6447368
“light”	108	0.3552632
Σύνολο	304*	1

*Το δείγμα μας έχει μέγεθος 308 αλλά υπάρχουν 4 NA's τιμές.

Από τον πίνακα διαπιστώνουμε ότι το 64% των διαμαντιών ανήκουν στη κατηγορία «heavy» και τα υπόλοιπα στην κατηγορία «light».

Θα εστιάσουμε τώρα την προσοχή μας στην κατηγορία των διαμαντιών τα οποία έχουν χρώμα “E”. Για την πιο εύκολη μελέτη αυτού θα δημιουργήσουμε ένα υποσύνολο με τον εξής τρόπο:

```
> E_colour=subset(a,colour=="E")
```

Μας ενδιαφέρει να περιγράψουμε την μεταβλητή price ξεχωριστά για κάθε κατηγορία της μεταβλητής best_diamond. Για λόγους απλότητας θα δημιουργήσουμε δυο επιπλέον υποσύνολα του προηγούμενου υποσυνόλου E_colour με ονόματα «H»(περιέχει μόνο «heavy» διαμάντια)και «L»(περιέχει μόνο «light» διαμάντια).

```

> H=subset(E_colour,best_diamond=="heavy")

> L=subset(E_colour,best_diamond=="light")

```

Θα περιγράψουμε με αριθμητικές μεθόδους τα παραπάνω υποσύνολα για την διεξαγωγή ορισμένων συμπερασμάτων για την τιμή (“price”). Με την εντολή fivenum() θα πάρουμε έναν ικανοποιητικό αριθμό στοιχείων για τα δείγματα:

```
> fivenum(H$price)
```

```
[1] 3346.0 5284.5 6689.5 10588.0 14051.0
```

Και

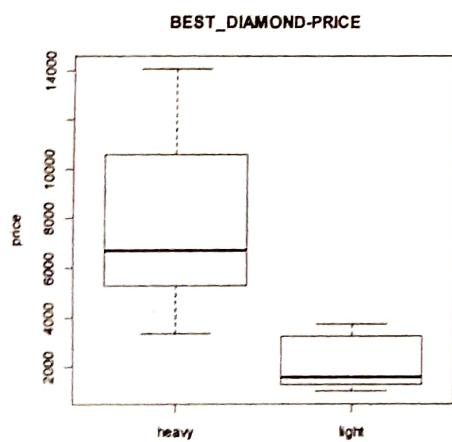
```
> fivenum(I$price)
```

```
[1] 1050.0 1287.5 1604.0 3221.5 3722.0
```

Για το Η δείγμα έχουμε τα εξής: 1)οι οριακές τιμές είναι 3334.6 και 14051.0, 2)το πρώτο τεταρτημόριο είναι 5284.5 ενώ το τρίτο 10588.0 και τέλος 3) η διάμεσος είναι 6689.5. Για το I δείγμα έχουμε τα εξής: 1)οι οριακές τιμές είναι 1050.0 και 3722.0, 2)το πρώτο τεταρτημόριο είναι 1287.5 ενώ το τρίτο 3221.5 και τέλος 3) η διάμεσος είναι 3722.0. Γρήγορα αντιλαμβανόμαστε από τα αποτελέσματα ότι η "Η" έχει πολύ μεγαλύτερη τιμή τόσο από τις ακραίες της τιμές (14051>3722) όσο και από το τρίτο τεταρτημόριο (10588>3221.5), οπότε τα διαμάντια με τύπο «heavy» είναι πολύ πιο ακριβά από τα «light».

Για να γίνει πιο κατανοητό θα κάνουμε ένα θηκόγραμμα για να αντιληφτούμε την υπεροχή αυτή της «heavy» όσο αναφορά την τιμή.

```
> plot(E_colour$best_diamond,E_colour$price,ylab="price",main="BEST_DIAMOND-PRICE")
```



Ερώτημα 5

Όπως και προηγούμενο, θα χρησιμοποιήσουμε για άλλη μια φορά την εντολή `ifelse()` προκειμένου να πάρουμε τις ζητούμενες πληροφορίες. Πιο συγκεκριμένα, θα φτιάξουμε μια επιπλέον στήλη στον a η οποία θα πάρει το όνομα `price_bin` και θα παίρνει ως τιμές τις: «**expensive**» αν το διαμάντι εχει τιμή μεγαλύτερη των 2500\$ και την τιμή «**cheap**» σε αντίθετη περίπτωση. Η αλληλουχία των εντολών είναι ως εξής:

```
> price_bin=ifelse(a[["price"]]>2500,"expensive","cheap")
```

```
> price_bin=factor(a$price_bin)
```

```
> a$price_bin=price_bin
```

Επίσης, με την χρήση των εντολών:

```
> table(price_bin)
```

```
> prop.table(table(price_bin))
```

Θα δημιουργήσουμε τον πίνακα συχνοτήτων και σχετικών συχνοτήτων:

	ΣΥΧΝΟΤΗΤΕΣ	ΣΧΣΤΙΚΕΣ ΣΥΧΝΟΤΗΤΣ
"expensive"	217	0.7068404
"cheap"	90	0.2931596
Σύνολο	307*	1

*Έχουμε ένα Να.

Στη συνέχεια, θα ασχοληθούμε με την σχέση μεταξύ των μεταβλητών «clarity» και «price_bin». Παρατηρούμε ότι και οι δυο μεταβλητές είναι κατηγορικές. Έχοντας αυτό το δεδομένο, θα χρειαστεί να υπολογίσουμε και να χρησιμοποιήσουμε ένα πίνακα συνάφειας. Για να δημιουργήσουμε ένα τέτοιο πίνακα θα εκτελέσουμε την παρακάτω εντολή:

```
> CPb=table(a$clarity,a$price_bin)
```

Ο πίνακας ο οποίος δημιουργείται τότε φαίνεται παρακάτω:

	Cheap	Expensive
IF	32	12
VS1	23	56
VS2	10	43
VVS1	10	42
VVS2	15	63

Ακολούθως, θα κατασκευάσουμε τις σχετικές συχνότητες κελιών, γραμμών και στηλών με τον παρακάτω τρόπο:

```
> prop.table(CPb)
```

```
> prop.table(CPb,1)
```

```
> prop.table(CPb,2)
```

Και σχεδιάζουμε τον πίνακα που περιέχει τις παραπάνω πληροφορίες:

	ΣΧ.ΣΥΧΝ.ΚΕΛΙΩΝ		ΣΧ.ΣΥΧΝ.ΓΡΑΜΜΩΝ		ΣΧ.ΣΥΧΝ.ΣΤΗΛΩΝ	
	Cheap	expensive	cheap	expensive	cheap	Expensive
IF	0.10457516	0.03921569	0.7272727	0.2727273	0.35555556	0.05555556
VS1	0.07516340	0.18300654	0.2911392	0.7088608	0.25555556	0.25925926
VS2	0.03267974	0.14052288	0.1886792	0.8113208	0.11111111	0.19907407
VSS1	0.03267974	0.13725490	0.1923077	0.8076923	0.11111111	0.19444444
VSS2	0.04901961	0.20588235	0.1923077	0.8076923	0.16666667	0.29166667

Παρατηρούμε αρχικά, ότι η κατηγορία της μεταβλητής IF έχει τα λιγότερα ακριβά διαμάντια(12) ενώ αντίστοιχα έχει και τα περισσότερα φθηνά. Η παραπάνω διαπίστωση είναι κοινή αν κοιτάξουμε και τον πίνακα σχετικών συχνοτήτων ανά γραμμή καθώς περιέχεται μια σχετικά μεγάλη συχνότητα (0,727...) για τα φθηνά διαμάντια και μια μικρή (0,272..) για τα ακριβά. Επίσης μια επιπλέον χαρακτηριστική τιμή είναι αυτή των 63 ακριβών διαμαντιών στη κατηγορία καθαρότερας VSS2.

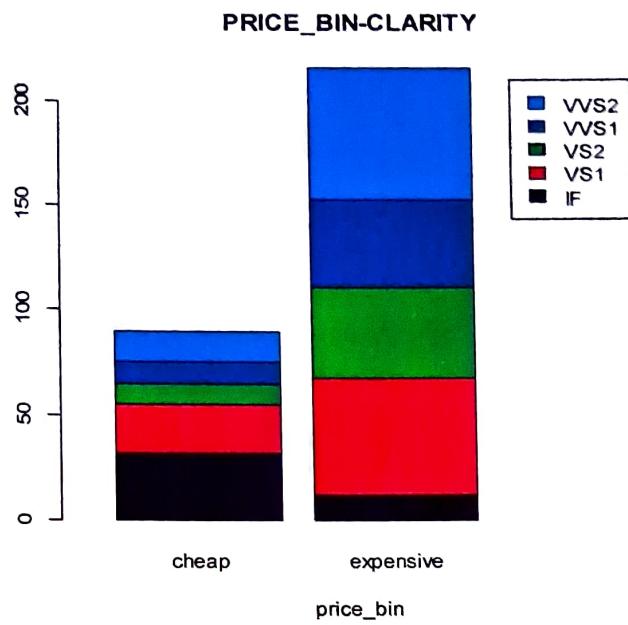
Σε σχέση τώρα με όλες τις υπόλοιπες κατηγορίες η «IF» διαθέτει το 36% περύπου από τα «cheap» διαμάντια ενώ μόνο το 5% από την ίδια κατηγορία βρίσκονται στα διαμάντια «expensive». Παράλληλα, η «VS1» περιέχει ένα ποσοστό 25% και από «cheap» αλλά και από «expensive» διαμάντια. Επιπλέον οι κατηγορίες «VS2» «VSS1» αποτελούν το 11% των «cheap» διαμαντιών και το 20% των «expensive». Τέλος, η κατηγορία «VSS2» αποτελεί το 16% των «cheap» και το 30% των «expensive» διαμαντιών.

Είμαστε σε θέση τώρα να δημιουργήσουμε ένα στοιβαγμένο και ένα ομαδοποιημένο ραβδόγραμμα με τα οποία θα δούμε πιο ξεκάθαρα τα παραπάνω χαρακτηριστικά γνωρίσματα του δείγματος.

Πρώτα θα κατασκευάσουμε το στοιβαγμένο ραβδόγραμμα ως εξής:

```
>barplot(CPb,xlab="price_bin",width=0.6,xlim=c(0,2),legend=levels(a$clarity),col=1:5
,main="PRICE_BIN-CLARITY")
```

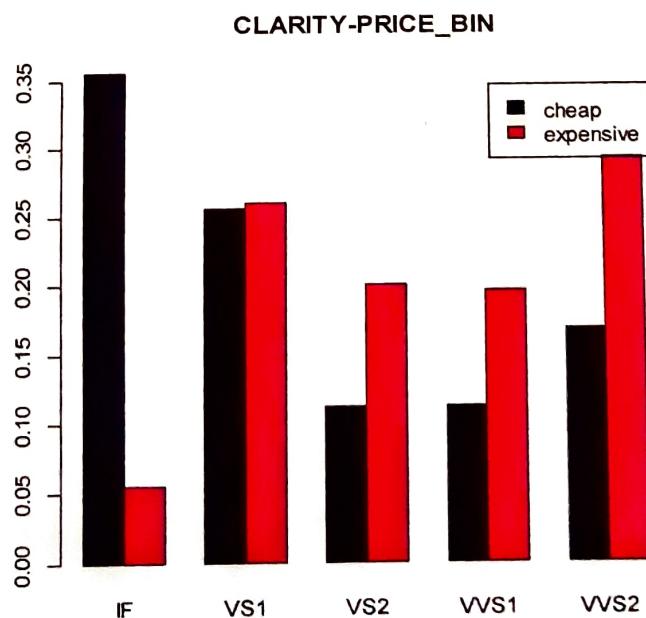
Και θα εμφανιστεί:



Σχεδιάζουμε και το ομαδοποιημένο ραβδόγραμμα :

```
>barplot(prop.table(table(a$price_bin,a$clarity),1),beside=T,legend=levels(a$price_b
in),col=c(1:2),main="CLARITY-PRICE_BIN")
```

Και παίρνουμε:



Άσκηση 2

Στην συγκεκριμένη άσκηση θα προσπαθήσουμε να επιλύσουμε το πρόβλημα αρχικών τιμών :

$$y'(t)=2*y(t)+t, \quad t \text{ ανήκει στο διάστημα } (0,2)$$

$$y(0)=y_0$$

με την βοήθεια της αριθμητικής μεθόδου Runge-Kutta.

Προσαρμόζουμε τον κώδικα που μας δίνεται για το πρόβλημά μας, ως εξής:

```
RK4<-function(n,y0)
```

```
{
```

```
y<-rep(0,n) #δημιουργώ ένα κενό διάνυσμα με n στοιχεία
```

```
y[1]<-y0 #αρχική τιμή του y
```

```
a<-0 #αριστερό άκρο διαστήματος
```

```
b<-2 #δεξιά άκρο διαστήματος
```

```
t<-a
```

```
h<-(b-a)/n
```

```
for(i in 2:n) #δομή επανάληψης για n-1 βήματα
```

```
{
```

```
k1=2*y[i-1]+t
```

```
k2=2*(y[i-1]+h*k1/2)+(t+h/2)
```

```
k3=2*(y[i-1]+h*k2/2)+(t+h/2)
```

```
k4=2*(y[i-1]+h*k3)+(t+h)
```

```
y[i]=y[i-1]+(h/6)*(k1+2*k2+2*k3+k4) #θέτω στην i-θέση την τιμή
```

```
t=t+h
```

```
}
```

```
return(y)
```

```
}
```

Με αυτόν τον τρόπο δημιουργούμε ένα διάνυσμα το οποίο περιέχει όλες τις τιμές που παίρνει του $y[i]$ σε κάθε βήμα ξεχωριστά. Καλούμε λοιπόν το διάνυσμα μας για $n=200$ και $y(0)=4$ και έτσι βρίσκουμε την λύση η οποία βρίσκεται στο τελευταίο στοιχείο του διανύσματος και είναι το 226.202394 ενώ καλώντας την συνάρτησή μας έχουμε το παρακάτω αποτέλεσμα (εμφανίζω κάποια στοιχεία από το διάνυσμα):

```
> RK4(200,4)
```

```
[1] 4.000000 4.080856 4.163446 4.247805 4.333970 4.421976
```

```
....
```

```
[97] 28.259073 28.839692 29.432140 30.036658 30.653488 31.282881
```

```
....
```

```
[163] 107.458317 109.645532 111.877033 114.153715 116.476489 118.846288
```

```
....
```

```
[199] 221.703634 226.202394
```