ΓΕΩΡΓΙΟΥ ΕΥΑΓΓΕΛΙΑ ΜΕΡΟΠΗ

ΕΙΣΑΓΩΓΗ ΣΤΗΝ R ΚΑΙ ΠΕΡΙΓΡΑΦΙΚΗ ΣΤΑΤΙΣΤΙΚΗ

Θέμα Έρευνας

Μια εταιρεία παραγωγής ηλεκτρονικών παιχνιδιών (video games), στα πλαίσια δοκιμής ενός παιχνιδιού ρόλων φαντασίας, εξάγει εσωτερικές δοκιμές στο παιχνίδι. Ενδιαφέρεται να ερευνήσει αν το παιχνίδι είναι εξίσου δίκαιο για όλων των ειδών χαρακτήρων που θα επιλέξουν οι παίκτες, ώστε να μην υπερτερεί σημαντικά κάποιος συνδυασμός των ιδιοτήτων που δίνονται στους χαρακτήρες. Για το σκοπό αυτό συνέλεξε πληροφορία από 48 δοκιμαστές, οι οποίοι δοκίμασαν το παιχνίδι σε μια συγκεκριμένη εικονική μάχη του παιχνιδιού με τον ίδιο (τεχνητής νοημοσύνης) αντίπαλο, για τις παρακάτω μεταβλητές: (α) ζημιά στον αντίπαλο (**damage**) σε κατάλληλη μονάδα μέτρησης που έχει θέσει η εταιρεία, (β) φύλο-ηλικία του χαρακτήρα (**sex)**, με κατηγορίες τις "man" (άντρας), "woman" (γυναίκα) και "children" (παιδί), (γ) φυλή του χαρακτήρα (**faction), με κατηγορίες τις "**human" (άνθρωποι), "ogre" (τέρατα) και "elf" (ξωτικά), (δ) κλάση του χαρακτήρα (**class)**, με κατηγορίες τις "wizard" (μάγος) και "warrior" (μαχητής) και (ε) συμπεριφορά (**attitude**), το οποίο είναι ένα ποσοτικό χαρακτηριστικό, η τιμή του οποίου υπολογίζεται συναρτήσει κάποιον αποφάσεων που έχουν παρθεί από τον παίχτη στη διάρκεια του παιχνιδιού.

Ανάλυση

Η ομάδα που διεξήγαγε τις δοκιμές μας παρέχει ένα αρχείο data.txt με τα ευρήματά της.
Αρχικά, ορίζουμε στην R το path για την φάκελο που είναι αποθηκευμένα τα data.txt και RScript μας με την εξής εντολή:

```
> setwd("C:/Users/user/Desktop/ΣΕΜΦΕ/6ο εξάμηνο/Ανάλυση Δεδομένων")
```

Η ομάδα που διεξήγαγε τις δοκιμές μας παρέχει ένα αρχείο data.txt με τα ευρήματά της. Το εισάγουμε στην R με τις εξής εντολές:

```
> rm(list=ls())
> Data <− read.table(file = "data.txt", header = T, na.strings = "uns")
```

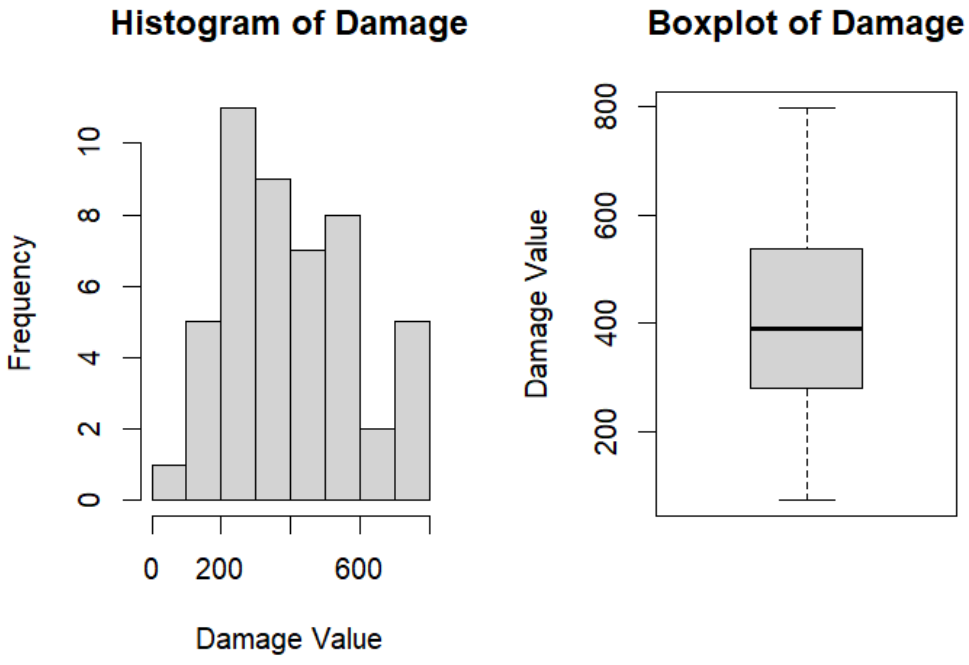
Με την παραπάνω εντολή αυτή, δηλώνουμε στην R την τοποθεσία του αρχείου .txt (όρισμα file), την ενημερώνουμε πως πράγματι η πρώτη γραμμή δηλώνονται ονόματα (header = T) και πως αντί του NA για τις αγνοούμενες τιμές έχει χρησιμοποιηθεί το “uns”. Η R επεξεργάζεται αυτή την εντολή και μας δίνει το πλαίσιο δεδομένων Data το οποίο είναι ένα data.frame και φαίνεται παρακάτω.

```
> Data
      damage      sex faction   class  attitude
1  277.48157      man   human  wizard  78.13099
2  556.91485 children   human  wizard 160.91966
3  341.55235    woman   human  wizard  96.39030
4  543.94284    woman    ogre warrior 155.81111
5  418.10003 children    ogre  wizard 117.76982
6  496.49359    woman    ogre  wizard 143.81055
7  727.44747    woman   human  wizard 214.06633
8  687.88652 children   human  wizard 201.52597
9  739.92653    woman    ogre  wizard 214.78265
10 297.00122    woman    ogre  wizard  82.21279
11  74.91456      man     elf  wizard  13.63325
12 383.35787    woman    ogre  wizard 110.81141
13 215.72197    woman   human  wizard  57.07777
14 513.07428      man     elf  wizard 147.39881
15 531.87218 children     elf  wizard 152.21807
16 513.82130      man    ogre  wizard 148.43328
17 311.94881    woman     elf warrior  83.80284
18 563.62944    woman     elf warrior 159.94769
19 431.35136 children    ogre warrior 124.46878
20 540.91727 children   human warrior 156.27282
21 446.61040      man    ogre warrior 127.76251
22 328.02945 children    ogre  wizard  92.16626
23 232.55587    woman     elf  wizard  61.81059
24 299.39612    woman    ogre warrior  84.86424
25 196.96243 children    ogre warrior  53.61770
26 608.12428    woman    ogre warrior 173.57687
27 289.07176    woman     elf warrior  79.90764
28 420.22759    woman    ogre warrior 119.47224
29 497.38974 children    ogre warrior 142.08133
30 162.62395    woman    ogre warrior  40.87574
31 280.67337    woman   human warrior  76.39351
32 399.27635      man    ogre  wizard 114.49457
33 721.87263 children    ogre  wizard 210.26646
34 377.40866      man     elf  wizard 105.00977
35 153.90149     <NA>    ogre warrior  39.52302
36 282.21515    woman    ogre  wizard  78.38405
37 398.26512      man    ogre warrior 111.06528
38 265.27044    woman     elf  wizard  71.36514
39 188.36572 children    ogre warrior  47.84708
40 364.54957 children    ogre     <NA> 101.05505
41 711.84000    woman     elf  wizard 208.99889
42 105.25878 children    ogre warrior  23.01371
43 427.83549 children     elf warrior 119.53519
44 282.46687 children   human  wizard  78.03080
45 378.89452 children   human warrior 107.93233
46 796.87913 children     elf  wizard 224.69405
47 551.79285      man    ogre warrior 159.24057
48 273.40465    woman   human warrior  74.86557
```

Damage:

Αρχικά, κάνουμε πρώτα ένα Ιστόγραμμα (Histogram) και ένα Θηκόγραμμα (Boxplot) για το damage. Αυτά γίνονται αντίστοιχα με τις εντολές:

```
> Damage <-  Data[,1]
> par(mfrow=c(1,2))
> hist(Damage, ylab="Frequency",xlab="Damage Value", main = "Histogram of Damage")
> boxplot(Damage, ylab="Damage Value", main="Boxplot of Damage")
title(main="Damage")
```



Από το Ιστόγραμμα φαίνεται πως

1. Οι περισσότεροι παίκτες έκαναν από 200 έως 600 damage.
2. Η κάπως ανομοιόμορφη κατανομή μεταξύ του 700 και 800 damage μας προϊδεάζει για το γεγονός πως μάλλον ο δειγματικός μέσος θα είναι μετατοπισμένος προς τα αριστερά.

Από το Θηκόγραμμα φαίνεται πως

1. Η ελάχιστη τιμή είναι κάτω από 100.
2. Το 25% των τιμών φαίνεται να είναι μεγαλύτερο από περίπου 300,
3. Η διάμεσος φαίνεται να είναι λίγο μικρότερη του 400 και
4. Το 75% των τιμών φαίνεται να είναι μεταξύ του 500 και του 550.
5. Η μεγαλύτερη τιμή του δείγματος φαίνεται να είναι λίγο μεγαλύτερη του 750.

Παρατηρούμε πως δεν υπάρχουν έκτροπες τιμές. Όλα τα παραπάνω θα επιβεβαιωθούν και με αριθμητικές μεθόδους οπότε δίνουμε στην R τις εντολές, οι οποίες θα δώσουν τα αντίστοιχα αποτελέσματα:

```
> fivenum(Damage)
[1]  74.91456 281.44426 390.81150 536.39472 796.87913
> summary(Damage)
Min.   1st Qu.  Median  Mean   3rd Qu.   Max.
74.91  281.83   390.81  408.51  534.13  796.88
> var(Damage)
> sd(Damage)
> IQR <− quantile(Damage,0.75)-quantile(Damage,0.25)
> Range <− max(Damage)-min(Damage)
```

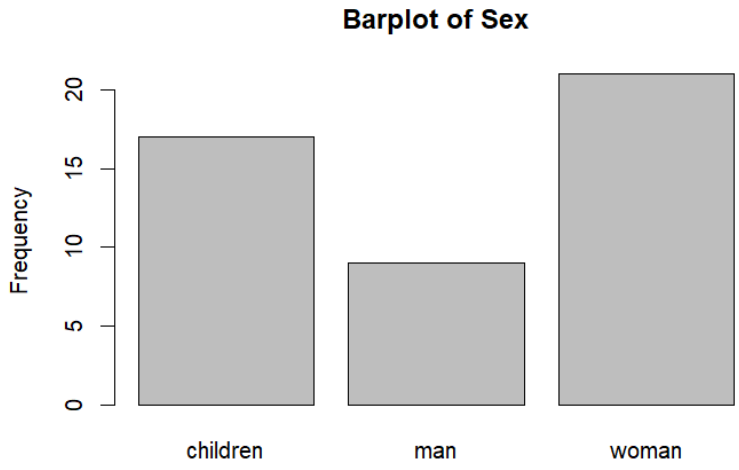
Διασπορά	Τυπική Απόκλιση	IQR	Range
31615.41	177.8072	252.3037	721.9646

Με την εντολή var() παίρνουμε την διασπορά, ενώ με την sd() την τυπική απόκλιση. Έχουμε, έτσι, μία εικόνα για το πόσο μακριά βρίσκονται οι παρατηρήσεις σε σχέση με τον δειγματικό μέσο. Το ενδοτεταρτημοριακό εύρος (IQR) μας δίνει και αυτό μία εικόνα για το μέτρο μεταβλητότητας των τιμών μας, χωρίς να επηρεάζεται από τις ακραίες τιμές. Στο τελευταίο κελί υπολογίσαμε το εύρος του δείγματος.

Sex:

Κάνουμε ένα Ραβδόγραμμα (Barplot) για το sex. Αυτά γίνονται αντίστοιχα με τις εντολές:

```
> Sex <− Data[,2]
> Sex <− as.factor(Sex)
> TSex <− table(Sex)
> par(mfrow=c(1,1))
> barplot(TSex, ylab= "Frequency",main="Barplot of Sex")
```



Από το Ραβδόγραμμα θα πάρουμε πληροφορίες σχετικά με το φύλο που επέλεξαν οι παίκτες και συγκεκριμένα φαίνεται πως οι περισσότεροι παίκτες προτίμησαν γυναίκες ή παιδιά σαν χαρακτήρες. Αναλυτικότερα, φαίνεται πως λιγότεροι από 10 παίκτες επέλεξαν άνδρα χαρακτήρα, λίγοι περισσότεροι από 15 επέλεξαν παιδί και λίγο περισσότεροι από 20 γυναίκα. Για μεγαλύτερη ακρίβεια, χρησιμοποιούμε αριθμητικές μεθόδους. Οπότε, δίνοντας τις παρακάτω εντολές στην R προκύπτουν αντίστοιχα:

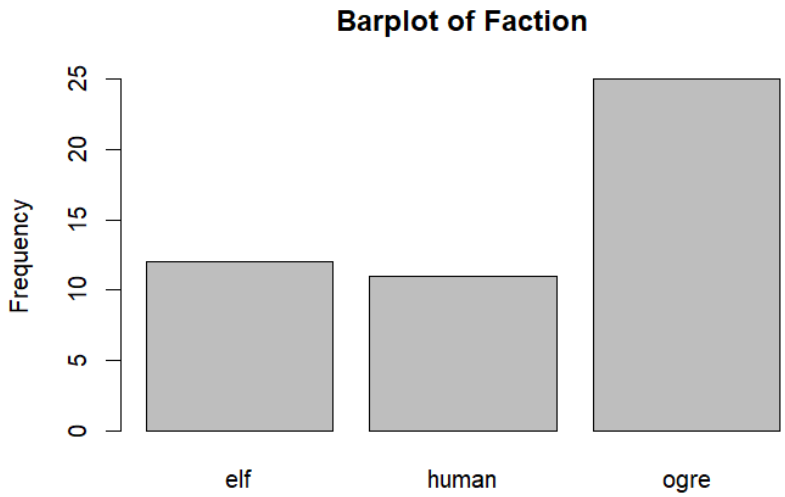
```
> table(Sex)
> prop.table(table(Sex))
```

Ακριβής Συχνότητα			Σχετική Συχνότητα		
children	man	woman	children	man	woman
17	9	21	0.3617021	0.1914894	0.4468085

Faction:

Κάνουμε αντίστοιχα ένα Ραβδόγραμμα (Barplot) για το faction. Αυτά γίνονται αντίστοιχα με τις εντολές:

```
> Faction <− Data[,3]
> Faction <− as.factor(Faction)
> TFaction <− table(Faction)
> par(mfrow=c(1,1))
> barplot(TFaction, ylab= "Frequency",main="Barplot of Faction")
```



Από το Ραβδόγραμμα θα πάρουμε πληροφορίες σχετικά με τη φυλή που επέλεξαν οι παίκτες και συγκεκριμένα φαίνεται πως οι περισσότεροι παίκτες προτίμησαν το τέρας, ενώ οι παίκτες που επέλεξαν να παίξουν ως ξωτικά είναι λίγο περισσότεροι από αυτούς που επέλεξαν να παίξουν ως άνθρωποι. Αναλυτικότερα, φαίνεται πως περίπου 25 παίκτες επέλεξαν το τέρας, περίπου12 επέλεξαν το ξωτικό και περίπου 11 τον άνθρωπο. Για μεγαλύτερη ακρίβεια, χρησιμοποιούμε αριθμητικές μεθόδους. Οπότε, δίνοντας τις παρακάτω εντολές στην R προκύπτουν αντίστοιχα:

```
> table(Faction)
> prop.table(table(Faction))
```

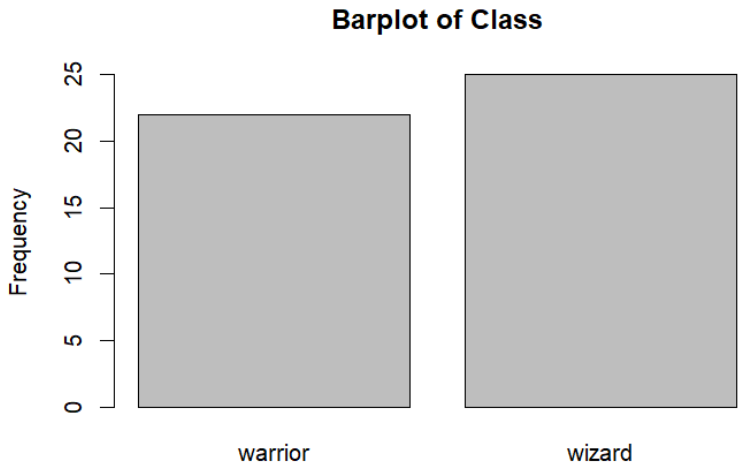
Ακριβής Συχνότητα			Σχετική Συχνότητα		
elf	human	ogre	elf	human	ogre
12	11	25	0.2500000	0.2291667	0.5208333

Class:

Ακολουθούμε ακριβώς την ίδια διαδικασία για το class

```
> Class <- Data[,4]
> Class <- as.factor(Class)
> TClass <- table(Class)
> par(mfrow=c(1,1))

> barplot(TClass, ylab= "Frequency",main="Barplot of Class")
```



Από το Ραβδόγραμμα θα πάρουμε πληροφορίες σχετικά με την κλάση που επέλεξαν οι παίκτες και συγκεκριμένα φαίνεται πως οι περισσότεροι παίκτες προτίμησαν τον μάγο και λιγότεροι τον μαχητοί. Αναλυτικότερα, φαίνεται πως περίπου 25 παίκτες επέλεξαν τον μάγο, ενώ περίπου 22 επέλεξαν το ξωτικό και περίπου 11 τον άνθρωπο. Για μεγαλύτερη ακρίβεια, χρησιμοποιούμε αριθμητικές μεθόδους. Οπότε, δίνοντας τις παρακάτω εντολές στην R προκύπτουν αντίστοιχα:

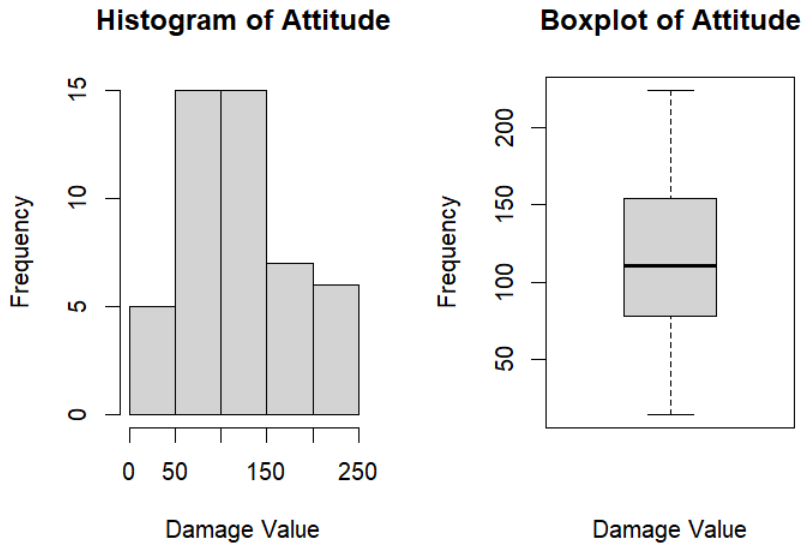
```
> table(Faction)
> prop.table(table(Faction))
```

Ακριβής Συχνότητα	Σχετική Συχνότητα
warrior wizard 22 25	warrior wizard 0.4680851 0.5319149

Attitude:

Ακολουθούμε αντίστοιχη διαδικασία με αυτή του damage οπότε προκύπτουν τα εξής:

```
> Attitude<-Data[,5]
> par(mfrow=c(1,2))
> hist(Attitude, ylab="Frequency",xlab="Damage Value", main = "Histogram of Attitude")
> boxplot(Attitude, ylab="Frequency",xlab="Damage Value", main = "Boxplot of Attitude")
```



Από το Ιστόγραμμα φαίνεται πως

1. Οι περισσότεροι παίκτες είχαν attitude από 50 έως 100.

Από το Θηκόγραμμα φαίνεται πως

1. Η ελάχιστη τιμή είναι κάτω από 25.
2. Το 25% των τιμών φαίνεται να είναι μεγαλύτερο από περίπου 75,
3. Η διάμεσος φαίνεται να είναι λίγο μεγαλύτερη του 100 και
4. Το 75% των τιμών φαίνεται να είναι περίπου ίσο με 150.
5. Η μεγαλύτερη τιμή του δείγματος φαίνεται να είναι λίγο μεγαλύτερη του 220.

Παρατηρούμε πως δεν υπάρχουν έκτροπες τιμές. Όλα τα παραπάνω θα επιβεβαιωθούν και με αριθμητικές μεθόδους οπότε δίνουμε στην R τις εντολές, οι οποίες θα δώσουν τα αντίστοιχα αποτελέσματα:

```
> fivenum(Attitude)
[1]  13.63325  78.08089 110.93834 154.01459 224.69405
> summary(Attitude)
Min. 1st Qu.  Median  Mean   3rd Qu.  Max.
13.63   78.11  110.94  115.57  153.12  224.69
> var(Attitude)
> sd(Attitude)
> IQR <- quantile(Attitude,0.75)-quantile(Attitude,0.25)
> Range <- max(Attitude)-min(Attitude)
```

Διασπορά	Τυπική Απόκλιση	IQR	Range
2859.23	53.47177	75.01039	211.0608

Με την εντολή var() παίρνουμε την διασπορά, ενώ με την sd() την τυπική απόκλιση. Έχουμε, έτσι, μία εικόνα για το πόσο μακριά βρίσκονται οι παρατηρήσεις σε σχέση με τον δειγματικό μέσο. Το ενδοτεταρτημοριακό εύρος (IQR) μας δίνει και αυτό μία εικόνα για το μέτρο μεταβλητότητας των τιμών μας, χωρίς να επηρεάζεται από τις ακραίες τιμές. Στο τελευταίο κελί υπολογίσαμε το εύρος του δείγματος.

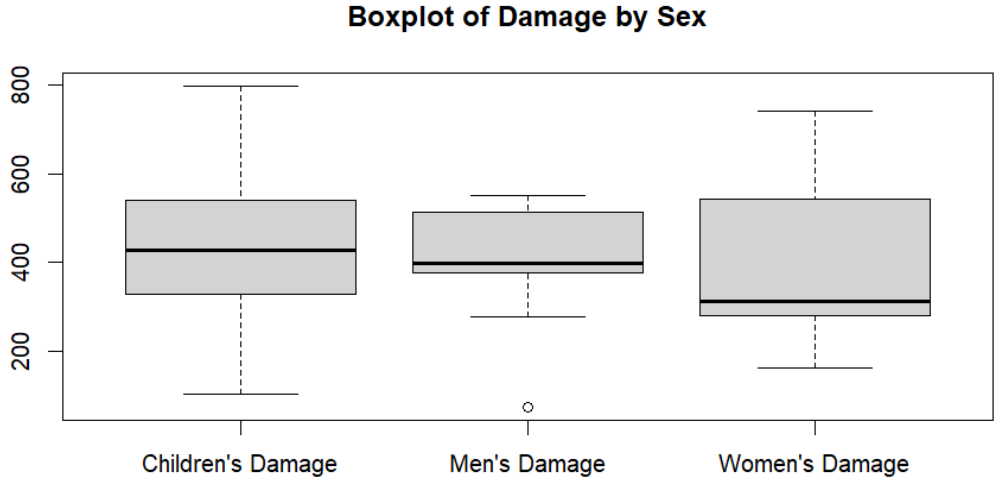
Σε συνέχεια της έρευνας θέλουμε να εξετάσουμε αν το το damage αλλάζει ανάλογα με τις υπόλοιπες μεταβλητές. Οπότε, δημιουργούμε τα αντίστοιχα subsets:

Damage-Sex:

```
> par(mfrow=c(1,1))
> Children <- subset(Data,sex == "children")
> Men <- subset(Data,sex == "man")
> Women <- subset(Data,sex == "woman")
```

Οπότε κάνουμε ένα Θηκόγραμμα (Boxplot) για το Damage του καθενός και έχουμε:

```
> boxplot(Children[,1],Men[,1],Women[,1],names=c("Children's
Damage","Men's Damage","Women's Damage"))
> title(main="Boxplot of Damage by Sex")
```



Αν και το τρίτο τεταρτημόριο είναι σχεδόν το ίδιο και για τα τρία, παρατηρούμε ότι η δειγματική διάμεσος του Damage των γυναικών είναι χαμηλότερη σε σχέση με των ανδρών και των παιδιών. Συμπεραίνουμε, λοιπόν, πως θα πρέπει να αυξηθεί το μέσο Damage που κάνουν οι γυναίκες, σε λογικά πλαίσια, για να μπορέσουμε να κρατήσουμε περίπου σταθερό το Q3. Επίσης,

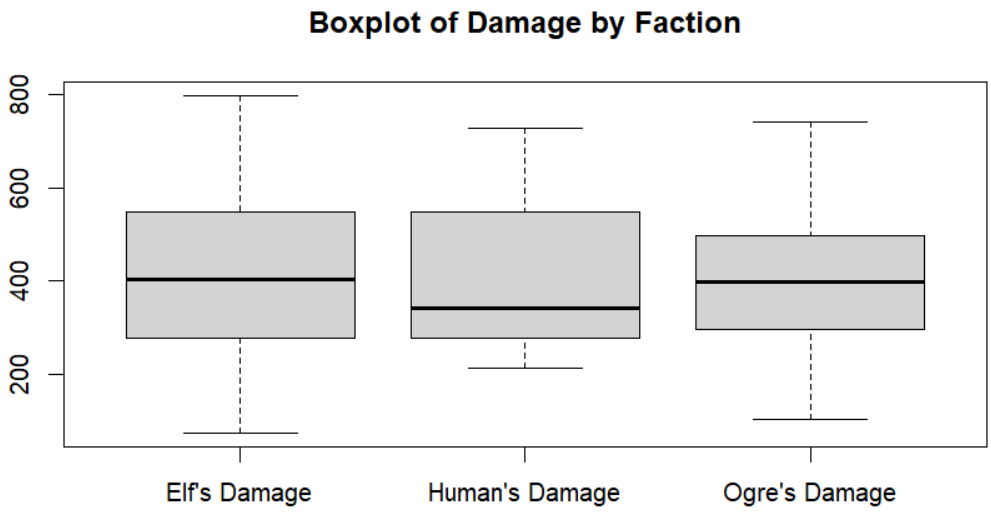
θα ήταν καλό να παρθούν μέτρα για να πάρουμε μεγαλύτερο δείγμα για τους άνδρες, γιατί στο δείγμα των μόλις 9 τιμών, η πιο μικρή θεωρήθηκε έκτροπη. Πιθανολογούμε ότι αυτό ευθύνεται και για την μεγάλη τιμή του Q1. Καταληκτικά, χρειάζεται να μετριάσουμε τις ακραίες τιμές των παιδιών, καθώς απέχουν αρκετά από τα Q3 και Q1 και των γυναικών, αφού απέχουν αρκετά από το Q3.

Damage-Faction:

```
> par(mfrow=c(1,1))
> Elf <- subset(Data,faction == "elf")
> Human <- subset(Data,faction == "human")
> Ogre <- subset(Data,faction == "ogre")
```

Οπότε κάνουμε ένα Θηκόγραμμα (Boxplot) για το Faction του καθενός και έχουμε:

```
> boxplot(Elf[,1],Human[,1],Ogre[,1],names=c("Elf's Damage","Human's Damage","Ogre's Damage"))
> title(main="Boxplot of Damage by Faction")
```



Σε σχεση με την προηγουμενη περιπτωση στο διαγραμμα φαινεται πως οι δειγματικές διάμεσοι είναι πολύ πιο κοντά σε αυτή την περίπτωση. Ιδανικά, θα θέλαμε να αυξήσουμε λίγο το μέσο Damage των ανθρώπων, κρατώντας όμως σταθερό το Q3. Κάτι ακόμα που μπορεί να διορθωθεί είναι οι ακραίες τιμές, που στην περίπτωση των ξωτικών και των τεράτων απέχουν πολύ από τα Q1 και Q3. Σκοπός μας θα πρέπει να είναι το Damage των χαρακτήρων να μην απέχει τόσο πολύ από τη διάμεσο.

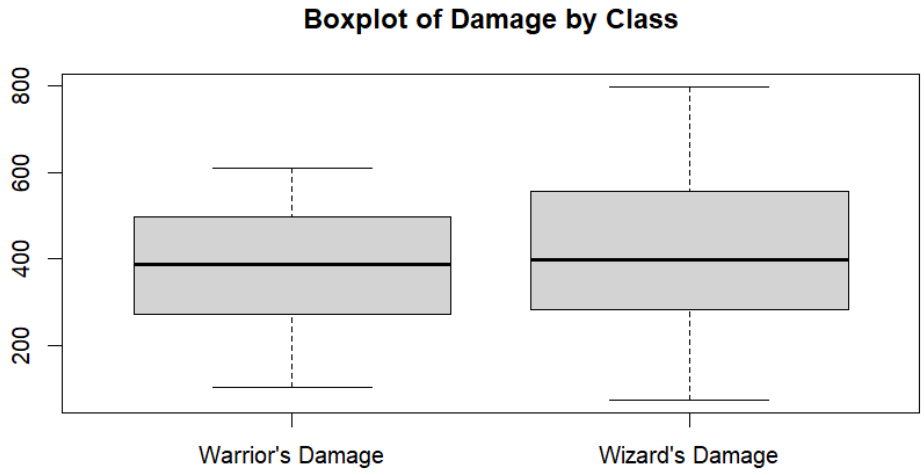
Damage-Class:

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

```
> Warrior <- subset(Data,class == "warrior")
> Wizard <- subset(Data,class == "wizard")
```

Οπότε κάνουμε ένα Θηκόγραμμα (Boxplot) για το Class του καθενός και έχουμε:

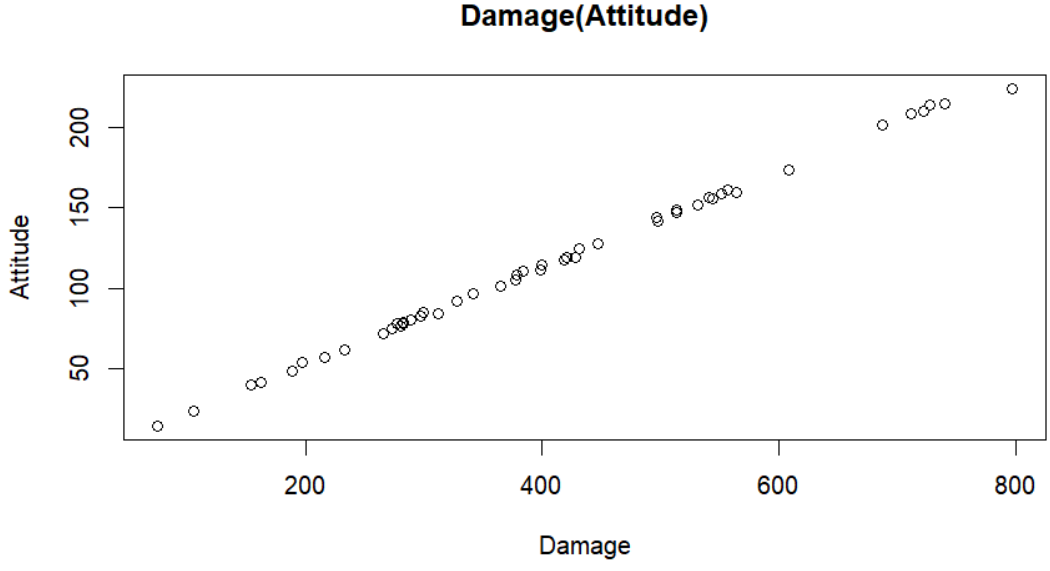
```
> boxplot(Warrior[,1],Wizard[,1],names=c("Warrior's Damage","Wizard's
Damage"))
 title(main="Boxplot of Damage by Class")
```



Όπως φαίνεται στο σχήμα, η δειγματική διάμεσος του Damage είναι περίπου ίδια για τους μάγους και τους μαχητές. Παρατηρούμε επίσης ότι το Q1 είναι επίσης στην ίδια περίπου τιμή. Τα Q3 διαφέρουν κατά περίπου 50 μονάδες, οπότε ίσως θα έπρεπε να ρίξουμε το Q3 του μάγου. Τέλος, αν μπορούσαμε να βελτιώσουμε κάτι άλλο, θα μειώναμε τις ακραίες τιμές και ειδικά αυτές του μάγου, στον οποίο απέχουν αρκετά από τα Q1 και Q3.

Damage-Attitude:

```
>plot(Damage,Attitude,xlab="Damage",ylab="Attitude",main="Damage(Attitu
de)")
```



Από το παραπάνω διάγραμμα παρατηρείται ξεκάθαρη γραμμική εξάρτηση του Damage από το Attitude.

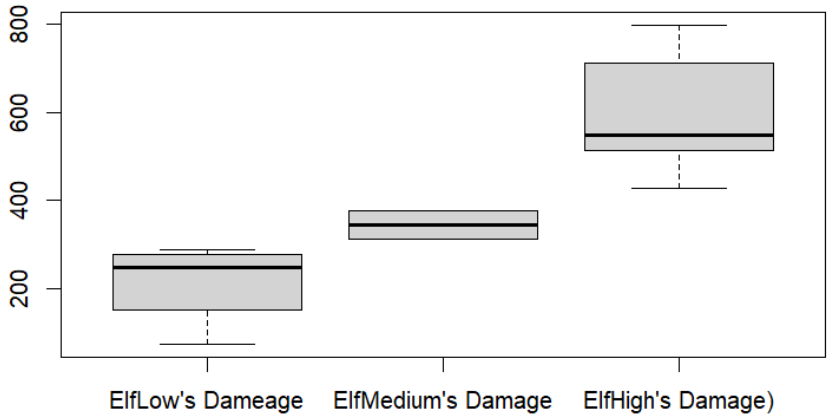
Στην συνέχεια, θα εξετάσουμε την τιμή Attitude. Αν είναι μικρότερη ή ίση του 80, του δίνουμε την κατηγορηματική μεταβλητή “Low”· αν ανήκει στο (80,110] την “Medium”, ενώ σε κάθε άλλη περίπτωση την “High”. Οπότε γράφουμε τον εξής κώδικα.

```
> n <− length(Attitude)
> f_attitude <−rep("High",48)
> f_attitude[Attitude<=80] <− "Low"
> f_attitude[Attitude>80 & Attitude<=110] <− "Medium"
> f_attitude<− as.factor(f_attitude)
> Data2 <− transform(Data,f_attitude=f_attitude)
> Tf_attitude <− table(f_attitude)
> Tf_attitude
> prop.table(Tf_attitude)
```

Ακριβής Συχνότητα			Σχετική Συχνότητα		
High	Low	Medium	High	Low	Medium
25	15	8	0.5208333	0.3125000	0.1666667

Και συνεχίζουμε φτιάχνοντας data frames, ένα για τις Low, ένα για τις Medium και ένα για τις High συμπεριφορές και τα αναπαριστούμε όλα σε ένα Θηκόγραμμα (Boxplot) ως εξής:

```
> Elf <− subset(Data2,faction == "elf")
> par(mfrow=c(1,1))
> ElfLow <− subset(Elf, f_attitude=="Low")
> ElfMedium <− subset(Elf, f_attitude == "Medium")
> ElfHigh <− subset(Elf, f_attitude == "High")
> boxplot(ElfLow[,1],ElfMedium[,1],ElfHigh[,1],names=c("ElfLow's
   Damage","ElfMedium's Damage","ElfHigh's Damage)"))
```



Παρ' ότι το δείγμα μας είναι πολύ μικρό γίνεται αντιπτό πως το Damage αυξάνεται ανάλογα με το Attitude. Ωστόσο, αν είχαμε μόνο τις κατηγορικές μεταβλητές, δεν θα μπορούσαμε να δούμε ότι η εξάρτησή τους είναι γραμμική.

Τέλος, όσον αφορά τον πίνακα συνάφειας μεταξύ των "τιμών" της μεταβλητής faction και των "τιμών" της μεταβλητής class, καθώς και το στοιβαγμένο και ομαδοποιημένο ραβδόγραμμα που τους αντιστοιχεί έχουμε τον εξής κώδικα:

```
> table(Faction,Class)
> prop.table(table(Faction,Class))
```

Πίνακας Συνάφειας			Σχετική συχνότητα κελιών		
	Class			Class	
Faction	warrior	wizard	Faction	warrior	wizard
elf	4	8	elf	0.08510638	0.17021277
human	4	7	human	0.08510638	0.14893617
ogre	14	10	ogre	0.29787234	0.21276596

```
> par(mfrow=c(1,2))
> barplot(FC_Table, xlim=c(0,3), xlab="Class", legend=levels(Faction),
col=1:3)
> barplot(FC_Table, xlim=c(0,10), xlab="Class", legend=levels(Faction),
col=1:3,
  beside=TRUE)
```

Είναι προφανές ότι τα wizard elf είναι διπλάσια από τα warrior elf, και τα wizard human είναι σχεδόν διπλάσια από τα warrior human. Ωστόσο, αυτή δεν είναι η κατάσταση με τα ogre, αφού όσοι παίζουν τέρας φαίνεται να προτιμούν την κλάση του μαχητή. Στο δείγμα μας, το πιο πιθανό είναι κάποιος να είναι warrior ogre με σχετική συχνότητα περίπου 0.297, ενώ τα λιγότερα πιθανά είναι τα human warrior ή elf warrior με μία σχετική συχνότητα περίπου 0.085.