Ανάλυση Δεδομένων με τη χρήση της R

Τελική Εργασία (Project)
Αξιολόγησης

Αλέξανδρος Νέζερης

14 Μαΐου 2017

1. Αρχικά, θα φορτώσουμε τα δεδομένα στην πλατφόρμα της R με την εντολή:
   > nbadata<-read.csv(file="NBA2016-data-for-final-project.csv",sep=";")
   Ενώ για να δούμε λίγα στατιστικά για τους πρώτους, κατά εμφάνιση παίκτες, εφαρμόζουμε την εντολή:
   > head(nbadata)

```
> nbadata<-read.csv(file="NBA2016-data-for-final-project.csv",sep=";")
> head(nbadata)
  Rk        Player Pos Age  Tm  G GS   MP  FG FGA X3P X3PA X2P X2PA  FT FTA ORB DRB TRB AST STL BLK TOV  PF PTS
1  1    Quincy Acy  PF  25 SAC 59 29  876 119 214  19   49 100  165  50  68  65 123 188  27  29  24  27 103 307
2  2  Jordan Adams  SG  21 MEM  2  0   15   2   6   0    1   2    5   3   5   0   2   2   3   3   0   2   2   7
3  3  Steven Adams   C  22 OKC 80 80 2014 261 426   0    0 261  426 114 196 219 314 533  62  42  89  84 223 636
4  4 Arron Afflalo  SG  30 NYK 71 57 2371 354 799  91  238 263  561 110 131  23 243 266 144  25  10  82 142 909
5  5 Alexis Ajinca   C  27 NOP 59 17  861 150 315   0    1 150  314  52  62  75 194 269  31  19  36  54 134 352
6  6  Cole Aldrich   C  27 LAC 60  5  800 134 225   0    0 134  225  60  84  86 202 288  50  47  68  64 139 328
```

2. Για να αφαιρέσουμε τις γραμμές που περιέχουν τουλάχιστον και ένα NA, χρησιμοποιούμε την εντολή:
   > nbadata<-na.omit(nbadata)

3. Για να υπολογίσουμε τον συνολικό αριθμό πόντων, εργαζόμαστε ως εξής:

```
> s<-rep(0,nlevels(nbadata$Tm))
> for(i in 1:dim(nbadata)[1]) {
+ for(j in 1:nlevels(nbadata$Tm)) {
+ if(as.character(nbadata[i,5])==levels(nbadata$Tm)[j]) {
+ s[j]=s[j]+nbadata[i,25] }
+ }
+ }
```

Αρχικοποίηση διανύσματος, στο οποίο θα εναποθέσουμε τους πόντους για κάθε ομάδα. Ανατρέχοντας στην στήλη που περιέχει τις ομάδες, βρίσκουμε τους παίκτες που ανήκουν στην ίδια ομάδα και αθροίζουμε τους πόντους τους .

Τα αποτελέσματα, που αντλήθηκαν, ήταν τα εξής:

Team	ATL	BOS	BRK	CHI	CHO	CLE	DAL	DEN	DET	GSW	HOU	IND	LAC
Points	8433	8669	7503	8335	8479	8554	8388	8355	8361	9421	8737	8377	8569

Team	LAL	MEM	MIA	MIL	MIN	NOP	NYK	OKC	ORL	PHI	PHO	POR	SAC
Points	7982	8126	8204	8122	8398	8423	8065	9038	8369	7142	8271	8622	8740

Team	LAL	MEM	MIA	MIL	MIN	NOP	NYK	OKC	ORL	PHI	PHO	POR	SAC
Points	7982	8126	8204	8122	8398	8423	8065	9038	8369	7142	8271	8622	8740

Team	SAS	TOR	UTA	WAS
Points	8490	8394	8010	8534

4. Για να βρούμε τον αριθμητικό μέσο πόντων για κάθε ομάδα, εργαζόμαστε ως εξής:

```
> s<-rep(0,nlevels(nbadata$Tm))
> n<-rep(0,nlevels(nbadata$Tm))
> for(i in 1:dim(nbadata)[1]) {
+ for(j in 1:nlevels(nbadata$Tm)) {
+ if(as.character(nbadata[i,5])==levels(nbadata$Tm)[j]){
+ s[j]=s[j]+nbadata[i,25]
+ n[j]=n[j]+1] }
+ }
+ }
> m = s/n
```

Αρχικοποίηση διανύσματος, στο οποίο θα εναποθέσουμε τους πόντους για κάθε ομάδα. Ανατρέχοντας στην στήλη που περιέχει τις ομάδες, βρίσκουμε τους παίκτες που ανήκουν στην ίδια ομάδα και α)αθροίζουμε τους πόντους τους, β)μετράμε πόσοι παίκτες υπάρχουν στην ομάδα. Τέλος, διαιρούμε τις 2 αυτές ποσότητες

Τα αποτελέσματα, που αντλήθηκαν, ήταν τα εξής:

Team	ATL	BOS	BRK	CHI	CHO	CLE	DAL	DEN
A.mean	496.0588	541.8125	468.9375	520.9375	498.7647	475.2222	524.25	439.7368

Team	DET	GSW	HOU	IND	LAC	LAL	MEM	MIA
A.mean	491.8235	588.8125	485.3889	523.5625	476.0556	532.1333	290.2143	431.7895

Team	MIL	MIN	NOP	NYK	OKC	ORL	PHI	PHO
A.mean	477.7647	524.875	401.0952	504.0625	531.6471	492.2941	420.1176	359.6087

Team	POR	SAC	SAS	TOR	UTA	WAS
A.mean	538.875	582.6667	499.4118	559.6	471.1765	449.1579

5. Για να προσδιορίσουμε τον συντελεστή μεταβλητότητας για κάθε ομάδα, εργαζόμαστε ως εξής:
   Αρχικά θα ορίσουμε μια συνάρτηση, μέσω της οποίας θα υπολογίζουμε τον συντελεστή μεταβλητότητας. Η συνάρτηση αυτή, θα είναι η εξής:

```
> cv<-function(x) {
+ z<-sd(x)*100/mean(x)
+ return(z)
+ }
```

```
> x<-rep(0,nlevels(nbadata$Tm))
> for(j in 1:nlevels(nbadata$Tm)) {
+ l = 0
+ for(i in 1:dim(nbadata)[1]) {
+ if(as.character(nbadata[i,5])==levels(nbadata$Tm)[j]){
+ l=c(l,nbadata[i,25]) }
+ }
+ z=cv(l[c(2:length(l))])
+ x[j]=z
+ }
```

Αρχικοποίηση διανύσματος, στο οποίο θα εναποθέσουμε τους συντελεστές μεταβλητότητας για κάθε ομάδα. Στη συνέχεια, ψάχνουμε τους παίχτες της κάθε ομάδας, και αποθηκεύομε σε ένα διάνυσμα τους πόντους και στο διάνυμα αυτό(αφαιρόντας το 1ο στοιχείο που είναι η αρχικοποίηση) εφαρμόζουμε τη συνάρτηση που ορίσαμε προηγουμένως. Τέλος, αποθηκεύουμε τους συντελεστές μεταβλητότητας σε ένα νέο διάνυσμα.

Τα αποτελέσματα, που αντλήθηκαν, ήταν τα εξής:

Team	ATL	BOS	BRK	CHI	CHO	CLE	DAL	DEN
CV	95.32819	93.97472	84.6899	80.50222	90.1876	109.1452	74.40185	87.43651

Team	DET	GSW	HOU	IND	LAC	LAL	MEM	MIA
CV	98.72945	111.0837	116.0387	93.5506	94.84117	79.29436	111.2367	102.1594

Team	MIL	MIN	NOP	NYK	OKC	ORL	PHI	PHO
CV	101.9777	103.0156	99.461	86.95525	118.1999	81.31848	78.5716	99.15181

Team	POR	SAC	SAS	TOR	UTA	WAS
CV	105.8698	84.11656	84.81179	96.60019	92.24712	95.02009

6. Οι εντολές, που θα χρησιμοποιήσουμε ώστε να εντωπίσουμε τις 5 πρώτες ομάδες με βάση τον αριθμό των παικτών, είναι οι εξής:

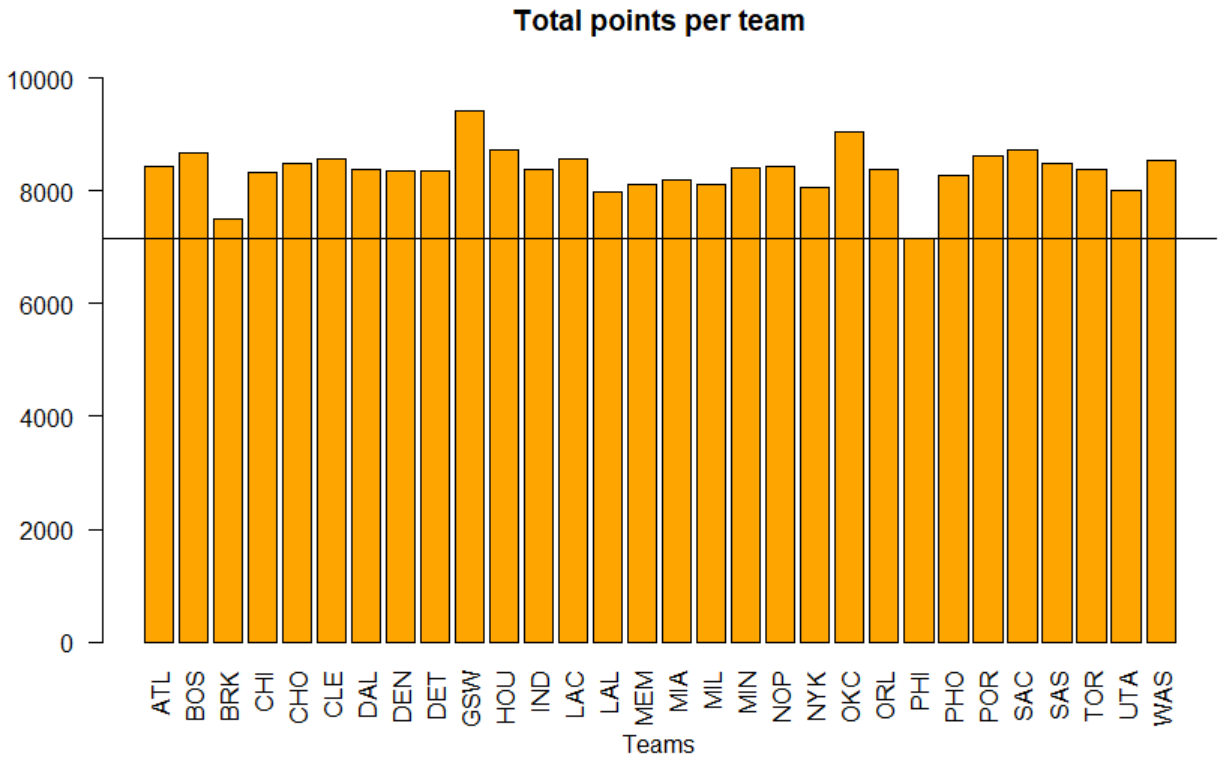
```
> n<-rep(0,nlevels(nbadata$Tm))
> for(j in 1:nlevels(nbadata$Tm)) {
+ for(i in 1:dim(nbadata)[1]) {
+   if(as.character(nbadata[i,5])==levels(nbadata$Tm)[j])
{
+ n[j]<-n[j]+1 }
+ } }
> g<-matrix(n,1)
> colnames(g)<-levels(nbadata$Tm)
> colnames(g)[order(g,decreasing=TRUE)][1:5])
> g[order(g,decreasing=TRUE)][1:5])
```

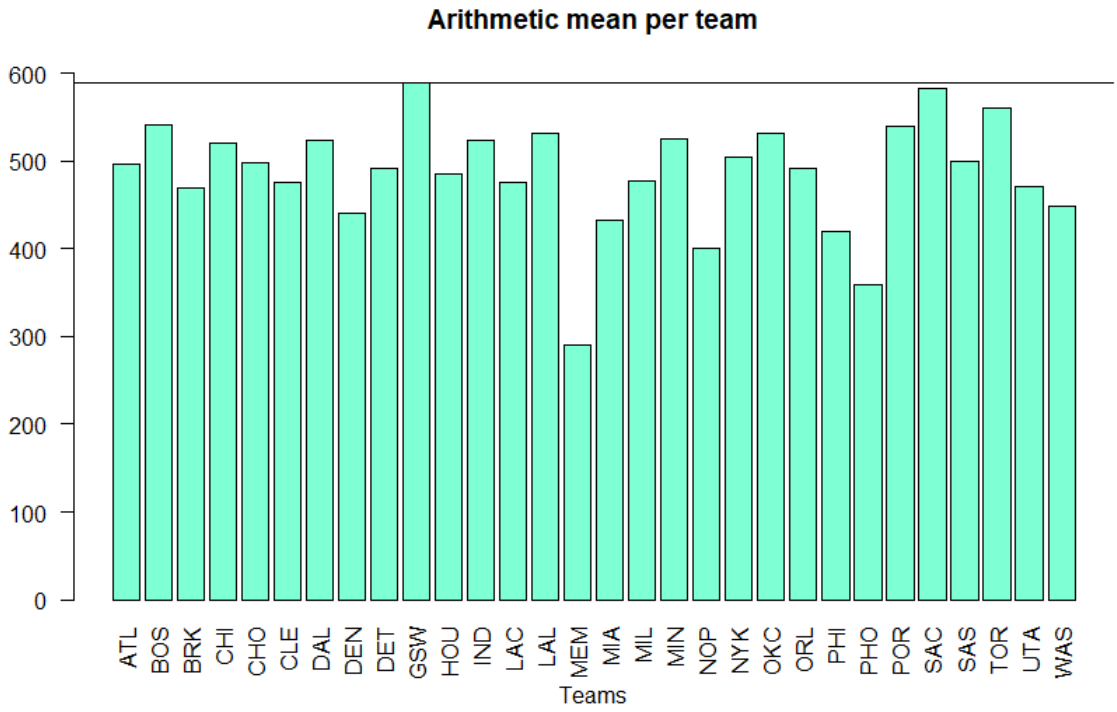
Ομοίως με πριν, θα δημιουργήσουμε ένα διάνυσμα, στο οποίο θα αναφέρεται ο αριθμός των παικτών ανά ομάδα. Έπειτα, θα μετατρέψουμε το διάνυσμα αυτό σε ένα πίνακα, ώστε να δώσουμε τα ονόματα των ομάδων σε κάθε περίπτωση. Στη συνέχεια, θα εκτυπώσουμε τα ονόματα των 5 ομάδων με τον μεγαλύτερο πλήθος παικτών.

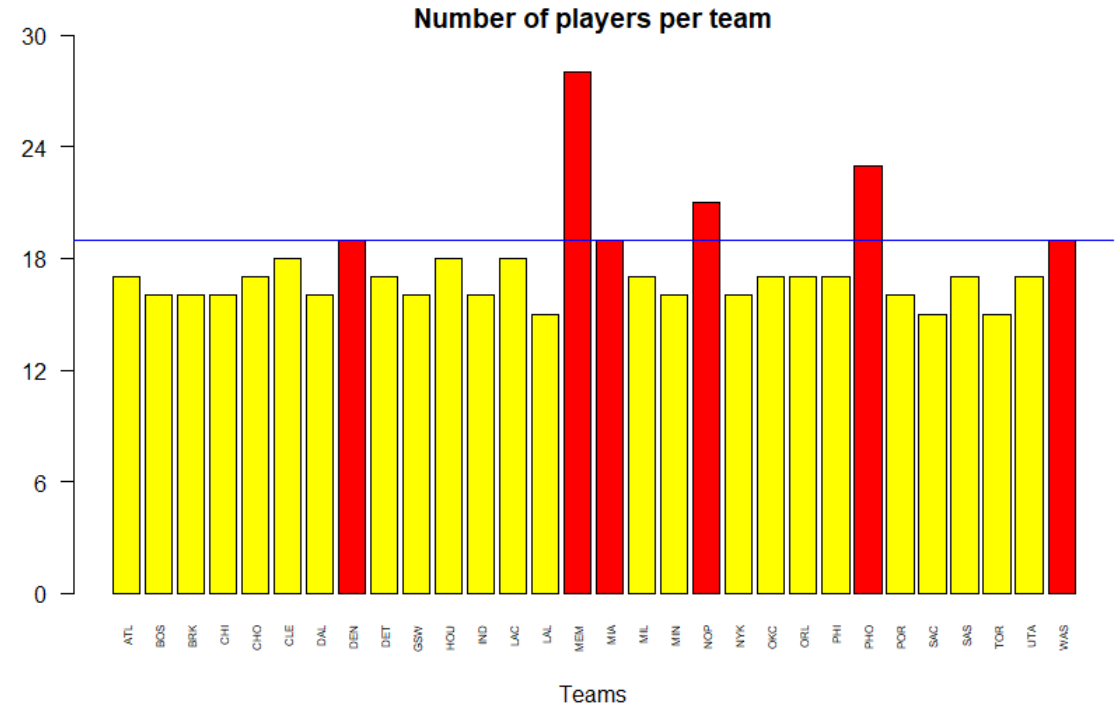
Από τα παραπάνω, έχουμε τα εξής:

Team	MEM	PHO	NOP	DEN	MIA
No. of Players	28	23	21	19	19

7. Με τη βοήθεια των πληροφορών που αντλήσαμε προηγουμένως, προκύπτουν τα εξής:







8. Καταρχάς, θα κατασκευάσουμε μία συνάρτηση, με την οποία θα βρίσκουμε τον συντελεστή Gini. Έτσι, λοιπόν, έχουμε:

```
> Gini<-function(x) {
+ s<-0
+ for(i in 1:length(x)) {
+ for(for(j in 1:length(x)) {
+ s< −s+abs(x[i]-x[j])
+ } }
+ g<-s/(2*length(x)*sum(x))
+ return(g)
```

Επομένως, εύκολα υπολογίζουμε τον συντελεστή Gini, με τον εξής τρόπο:

```
> gp1<-NULL
> gp2<-NULL
> for(j in 1:nlevels(nbadata$Tm)) {
+ e1<-NULL
+ e2<-NULL
+ for(i in 1:dim(nbadata)[1]) {
+ if(as.character(nbadata[i,5])==levels(nbadata$Tm)[j]) {
+ for(e1<-c(gp1,nbadata$PTS[i])
+ for(e2<-c(gp1,nbadata$MP[i])
+ gp1[j]<-Gini(e1)
+ gm1[j]<-Gini(e2)
+ } } }
> names(gp1)<-levels(nbadata$Tm)
> names(gm1)<-levels(nbadata$Tm)
> gp1
> gm1
```

Τα αποτελέσματα που προέκυψαν, με στρογγυλοποίηση σε 4 δεκαδικά ψηφία, ως προς τους πόντους είναι τα εξής:

Team	ATL	BOS	BRK	CHI	CHO	CLE	DAL	DEN
Gini co.	0.4996	0.6637	0.7484	0.7926	0.8282	0.8453	0.8697	0.8198

Team	DET	GSW	HOU	IND	LAC	LAL	MEM	MIA
Gini co.	0.8834	0.7980	0.8998	0.8718	0.8213	0.9071	0.8252	0.9174

Team	MIL	MIN	NOP	NYK	OKC	ORL	PHI	PHO
Gini co.	0.8878	0.9391	0.8970	0.9318	0.9458	0.8476	0.7485	0.5751

Team	POR	SAC	SAS	TOR	UTA	WAS
Gini co.	0.8985	0.9412	0.9277	0.7915	0.8748	0.9526

Ενώ για τα λεπτά παιχνιδιού κάθε ομάδας έχουμε:

Team	ATL	BOS	BRK	CHI	CHO	CLE	DAL	DEN
Gini co.	0.4998	0.6651	0.7493	0.7981	0.8315	0.8518	0.8727	0.8707

Team	DET	GSW	HOU	IND	LAC	LAL	MEM	MIA
Gini co.	0.8952	0.8714	0.9076	0.9018	0.8802	0.9232	0.8900	0.9357

Team	MIL	MIN	NOP	NYK	OKC	ORL	PHI	PHO
Gini co.	0.9316	0.9425	0.9287	0.9416	0.9486	0.9306	0.8475	0.7457

Team	POR	SAC	SAS	TOR	UTA	WAS
Gini co.	0.9455	0.9555	0.9495	0.8784	0.9342	0.9594

9. Αρχικά, θα εντοπίσουμε τους παίκτες που εμφανίζονται τα στοιχεία τους τουλάχιστον 2 φορές, καθώς έπαιξαν σε διαφορετικές ομάδες μέσα στη χρονιά, και θα ενοποιήσουμε τα στοιχεία τους. Έπειτα, θα αφαιρέσουμε τους παίκτες που έπαιξαν κατά μέσο όρο λιγότερο από 5 λεπτά και τέλος θα εφαρμοσουμε τη διαδικασία που ακολουθήσαμε στο προηγούμενο ερώτημα.

```
> nbanewdata<-nbadata
> po<-NULL
> for(i in 1:(dim(nbanewdata)[1]-1)) {
+ for(j in (i+1):dim(nbanewdata)[1]) {
+ if(as.character((nbanewdata$Player[i])==as.character(nbanewdata$Player[j])) {
+ po<-c(po,j)
+ } } }
> for(i in 1:(length(po))) {
+ if(po[i]==136|po[i]==230) {
+ print(i))
+ }
+ }
> po2<-pos[-c(8,9,23,24)]
> inpo<-po2-1)
> nbanewdata[inpo,6:25]<-nbanewdata[inpo,6:25]+nbanewdata[po2,6:25]
> nbanewdata[134,6:25]<-nbanewdata[134,6:25]+nbanewdata[136,6:25]
> nbanewdata[228,6:25]<-newdata[228,6:25]+nbanewdata[232,6:25]
> nbanewdata2<-nbanewdata[-unique(pos),]
> v5<-NULL)
> for(i in 1:dim(nbanewdata2)[1]) {
+ if(nbanewdata2$MP[i]/nbanewdata2$G¡5) {
+ v5<-c(v5,i)
+ } }
> gininbadata<-nbanewdata[-v5,]
> gp2<-NULL
```

```
> gm2<-NULL
> for(j in 1:nlevels(gininbadata$Tm)) {
+ e3<-NULL
+ e4<-NULL
+ for(i in 1:dim(gininbadata)[1]) {
+ if(as.character(gininbadata[i,5])==levels(gininbadata$Tm)[j]) {
+ for(e3<-c(gp2,gininbadata$PTS[i])
+ for(e4<-c(gp2,gininbadata$MP[i])
+ gp2[j]<-Gini(e3)
+ gm2[j]<-Gini(e4)
+ } } }
> names(gp2)<-levels(nbadata$Tm)
> names(gm2)<-levels(nbadata$Tm)
> gp2
> gm2
> names(GP)<-levels(nbadata$Tm)
> names(GM)<-levels(nbadata$Tm)
> GP
> GM
```

Τα αποτελέσματα που προέκυψαν, με στρογγυλοποίηση σε 4 δεκαδικά ψηφία, ως προς τους πόντους είναι τα εξής:

Team	ATL	BOS	BRK	CHI	CHO	CLE	DAL	DEN
Gini co.	0.9491	0.9071	0.9468	0.9027	0.9321	0.9027	0.9427	0.7343

Team	DET	GSW	HOU	IND	LAC	LAL	MEM	MIA
Gini co.	0.9099	0.9546	0.9294	0.8360	0.7197	0.9112	0.7495	0.9230

Team	MIL	MIN	NOP	NYK	OKC	ORL	PHI	PHO
Gini co.	0.8703	0.9539	0.9441	0.9371	0.9553	0.8265	0.7116	0.5318

Team	POR	SAC	SAS	TOR	UTA	WAS
Gini co.	0.8911	0.9410	0.9270	0.7835	0.8703	0.9526

Ενώ για τα λεπτά παιχνιδιού κάθε ομάδας έχουμε:

Team	ATL	BOS	BRK	CHI	CHO	CLE	DAL	DEN
Gini co.	0.9574	0.9356	0.9580	0.9499	0.9546	0.9372	0.9567	0.8962

Team	DET	GSW	HOU	IND	LAC	LAL	MEM	MIA
Gini co.	0.9505	0.9590	0.9469	0.9100	0.8449	0.9454	0.8682	0.9573

Team	MIL	MIN	NOP	NYK	OKC	ORL	PHI	PHO
Gini co.	0.9447	0.9595	0.9573	0.9516	0.9592	0.9331	0.8282	0.7143

Team	POR	SAC	SAS	TOR	UTA	WAS
Gini co.	0.9481	0.9585	0.9512	0.8749	0.9332	0.9593

Αφαιρώντας από τους συντελεστές Gini που βρήκαμε στην ερώτηση 8, τους συντελεστές που βρήκαμε σε αυτό το ερώτημα, μπορούμε να πρατηρήσουμε αν οι συντελεστές αυξήθηκαν ή μειώθηκαν.
Ως προς τους πόντους κάθε ομάδας, η μεταβολή στους συντελεστές Gini, είναι η εξής:

Team	ATL	BOS	BRK	CHI	CHO	CLE	DAL	DEN
Gini co. dif. per pts	-0.4495	-0.2434	-0.1984	-0.1101	-0.1039	-0.0574	-0.0730	0.0855

Team	DET	GSW	HOU	IND	LAC	LAL	MEM	MIA
Gini co. dif. per pts	-0.0265	-0.1566	-0.0295	0.0358	0.1016	-0.0040	0.0758	-0.0056

Team	MIL	MIN	NOP	NYK	OKC	ORL	PHI	PHO
Gini co. dif. per pts	0.0175	-0.0148	-0.0471	-0.0053	-0.0095	0.0211	0.0369	0.0432

Team	POR	SAC	SAS	TOR	UTA	WAS
Gini co. dif. per pts	0.0074	0.0001	0.0007	0.0080	0.0045	0.0001

Ενώ για τα λεπτά παιχνιδιού κάθε ομάδας έχουμε:

Team	ATL	BOS	BRK	CHI	CHO	CLE	DAL	DEN
Gini co. dpm	-0.4576	-0.2705	-0.2088	-0.1519	-0.1231	-0.0854	-0.0840	-0.0256

Team	DET	GSW	HOU	IND	LAC	LAL	MEM	MIA
Gini co. dpm	-0.0553	-0.0876	-0.0393	-0.0082	0.0353	-0.0222	0.0218	-0.0216

Team	MIL	MIN	NOP	NYK	OKC	ORL	PHI	PHO
Gini co. dpm	-0.0132	-0.0170	-0.0286	-0.0100	-0.0107	-0.0025	0.0193	0.0314

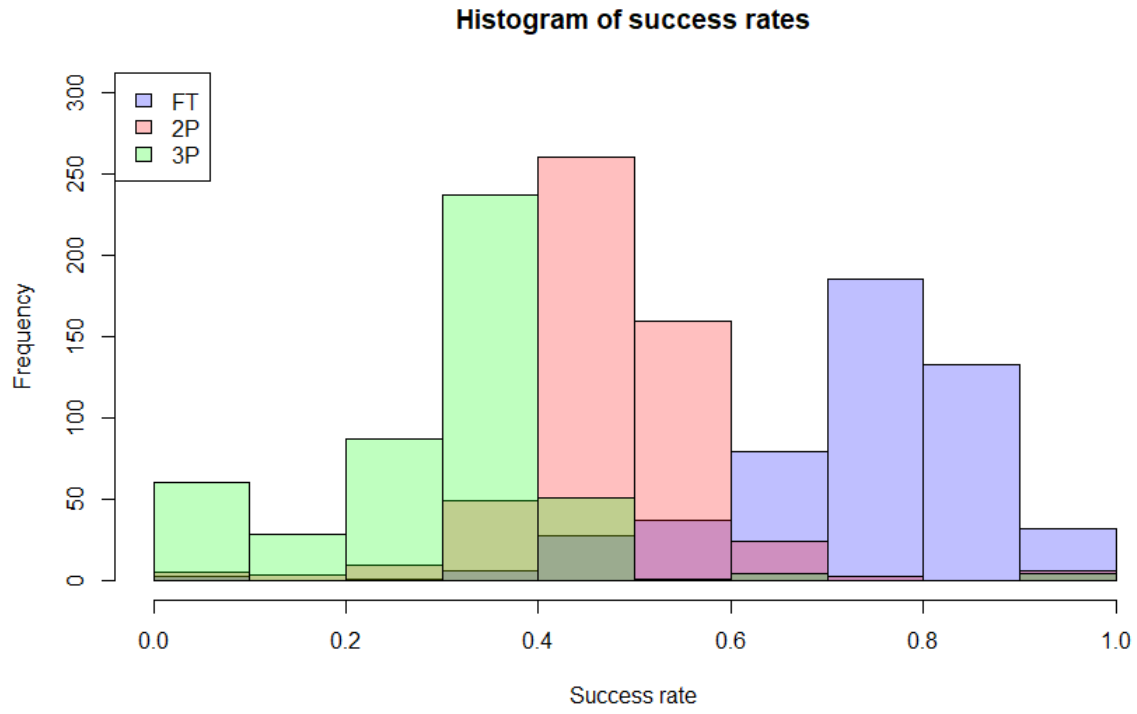
Team	POR	SAC	SAS	TOR	UTA	WAS
Gini co. dpm	-0.0025	-0.0031	-0.0017	0.0035	0.0010	0.0000

10. Το γράφημα με τα ιστογράμματα για τα ποσοστά ευστοχίας θα προκύψει από τις εξής εντολές:

```
> h1<-hist(nbadata$FT/nbadata$FTA)
```

```
> h2<-hist(nbadata$X2P/nbadata$X2PA)
> h3<-hist(nbadata$X3P/nbadata$X3PA)
> plot( h1, col=rgb(0,0,1,1/4),ylim=c(0,300),main="Histogram of success rates",xlab="Success rate")
> plot( h2, col=rgb(1,0,0,1/4),ylim=c(0,300), add=T)
> plot( h3, col=rgb(0,1,0,1/4),ylim=c(0,300), add=T)
> legend('topleft',c('FT','2P','3P'),fill=c(rgb(0,0,1,1/4),rgb(1,0,0,1/4),rgb(0,1,0,1/4))))
```

Από τα παραπάνω, προκύπτει το εξής γράφημα:

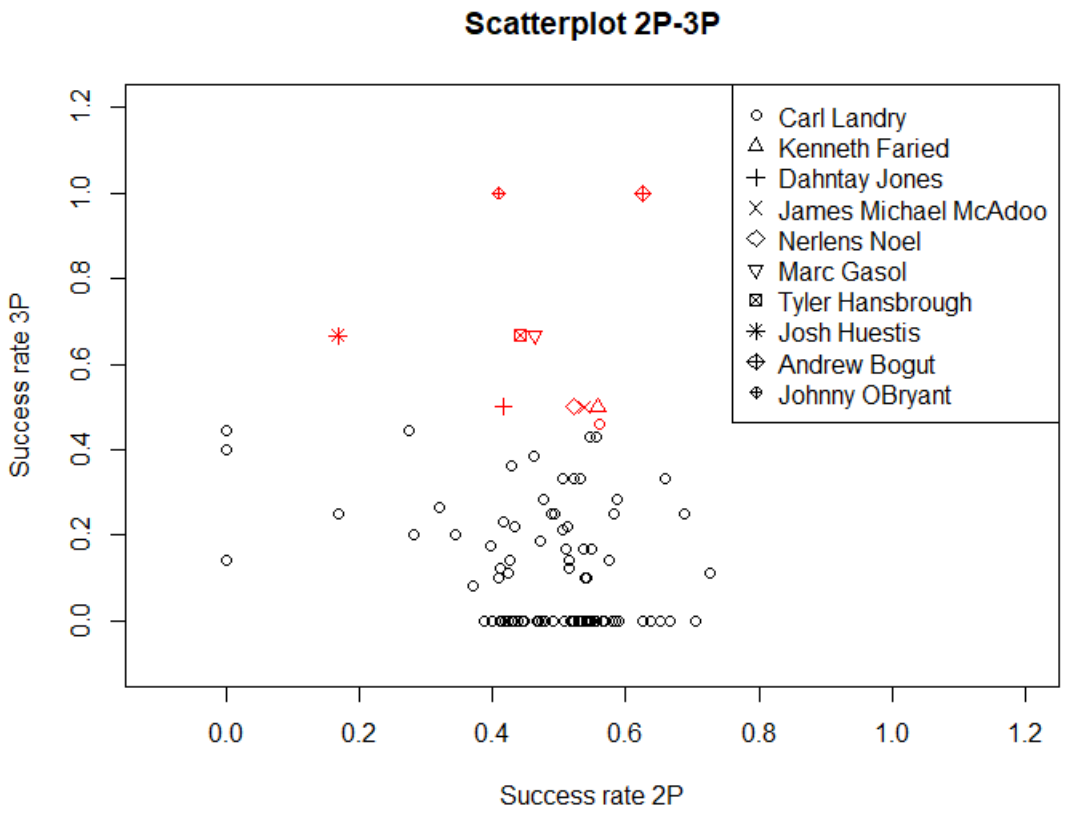


11. Αρχικά, θα εντοπίσουμε τους παίκτες που εμφανίζονται από 2 φορές και πάνω, με σκοπό να προσθέσουμε τα δεδομένα τους, και θα κρατήσουμε μόνο τα στοιχεία των παικτών αυτών από μια φορα. Στη συνέχεια, θα αφαιρέσουμε από τα δεδομένα μας, του παίκτες που έπαιξαν λιγότερο από 5 λεπτά καθώς και αυτούς που πραγματοποίητσαν περισσότερα από 20 σουτ 3 πόντων. Έπειτα, θα εισάγουμε ένα νέο διάνυσμα, το οποίο θα αναδυκνύει την ευστοχία των παικτών στα τρίποντα, και με τη βοήθεια αυτού του διανύσματος, θα τους κατατάξουμε κατά αύξουσα σειρά. Τέλος, θα παραστήσουμε στο γράφημά μας τους 10 πιο έυστοχους παίκτες, χρησιμοποιόντας κόκκινο χρώμα αλλά και εισάγοντας τα ονόματά τους. Τέλος, θα προσθέσουμε και τα σημεία που αντίστοιχουν στους υπόλοιπους παίκτες. Όλα αυτά θα πραγματοποιηθούν με τις εξής εντολές:

```
> newdata<-nbadata
> pos<-NULL
> for(i in 1:(dim(newdata)[1]-1)) {
```

```
+ for(j in (i+1):dim(newdata)[1]) {
+ if(as.character((newdata$Player[i])==as.character(newdata$Player[j])) {
+ pos<-c(pos,j)
+ } } }
> for(i in 1:(length(pos))) {
+ if(pos[i]==136|pos[i]==230) {
+ print(i))
+ }
+ }
> pos2<-pos[-c(8,9,23,24)]
> inpos<-pos2-1)
> newdata[inpos,6:25]<-newdata[inpos,6:25]+newdata[pos2,6:25]
> newdata[134,6:25]<-newdata[134,6:25]+newdata[136,6:25]
> newdata[228,6:25]<-newdata[228,6:25]+newdata[232,6:25]
> newdata2<-newdata[-unique(pos),]
> vector1<-NULL)
> vector2<-NULL)
> for(i in 1:dim(newdata2)[1]) {
+ if(newdata2$MP[i]/newdata2$G¡5) {
+ vector1¡-c(vector1,i)
+ } }
> for(j in 1:dim(newdata2)[1]) {
+ if(newdata2$X3PA[j]¿=20) {
+ vector2¡-c(vector2,j)
+ } }
> newdata3<-newdata2[-c(vector1,vector2),]
> vector3<-newdata3$X3P/newdata3$X3PA
> vector3[is.nan(vector3)]<-0
> newdata4<-cbind(newdata3,vector3)
> library(plyr)
> newdata5<-arrange(newdata4,vector3)
> st<-dim(newdata5)[1]-9
> en<-dim(newdata5)[1]
> vector4<-NULL
> for(i in st:en) {
+ for(j in 1:dim(newdata3)[1]) {
+ if(as.character(newdata5Player[i]) == as.character(newdata3Player[j])) {
+ vector4<c(vector4,j)
+ } } }
> plot(newdata3$X2P[vector4]/newdata3$X2PA[vector4],
+ newdata3$X3P[vector4]/newdata3$X3PA[vector4],pch=c(1:10),
+xlim=c(-0.1,1.2),ylim=c(-0.1,1.2),col=2, main='Scatterplot 2P-3P',
```

```
+xlab='Success rate 2P',ylab='Success rate 3P')
> list<-newdata3$Player[vector4]
> legend("topright", legend=list, pch=c(1:10))
> points(newdata3$X2P[-vector4]/newdata3$X2PA[-vector4],
+ newdata3$X3P[-vector4]/newdata3$X3PA[-vector4])
```



12. , 13. Αρχικά, θα βρούμε το level, όπου βρίσκεται η θέση PG, με τον εξής τρόπο:

```
> levels(nbadata$Pos)
[1] "C" "PF" "PG" "SF" "SG"
```

Επομένως, η θέση PG βρίσκεται στο επίπεδο 3 των θέσεων. Ο αλγόριθμος που θα χρησιμοποιήσουμε ώστε να εντοπίσουμε και να εμφανίσουμε τους παίκτες με τα περισσότερα rebounds, μαζί μετα στατιστικά τους, είναι ο εξής:
Θα αρχικοποιήσουμε 3 διανύσματα h,j,int, στα οποία στη συνέχεια θα εναποθέτουμε τα rebounds, τα ονόματα των παικτών και τον αύξοντα αριθμό με τη σειρά που μας παρουσιάζονται αντίστοιχα. Έπειτα, συγκρίνουμε τον αριθμό των rebounds και εμφανίζουμε στην οθόνη τα ζητούμενα:

```
> h<-NULL
> j<-NULL
```

```
> int<-NULL
> for(i in 1:dim(nbadata)[1]) {
+ if(as.character(nbadata[i,3])==levels(nbadata$Pos)[3]) {
+ h<c(h,nbadata[i,19])
+ j<-c(j,as.character(nbadata[i,2]))
+ int<-c(int,i) }
+ }
> k<-matrix(h,1)
> colnames(k)<-levels(nbadata$Pos)
> colnames(k)[order(k,decreasing=TRUE)][1:5])
> k[order(k,decreasing=TRUE)][1:5])
> l<-matrix(h,1)
> colnames(l)<-int
> ve<-as.numeric(colnames(l)[order(l,decreasing=TRUE)][1:5])
> for(i in 1:length(ve)) {
+ print(nbadata[ve[i],])
+ }
```

Players	Russell Westbrook	Giannis Antetokounmpo	Rajon Rondo	Stephen Curry	John Wall
TRB	626	612	435	430	379

```
     Rk            Player Pos Age  Tm  G GS   MP  FG  FGA X3P X3PA X2P X2PA  FT FTA ORB DRB TRB AST STL BLK TOV  PF  PTS
504 452 Russell Westbrook  PG  27 OKC 80 80 2749 656 1444 101  341 555 1103 465 573 145 481 626 834 163  20 342 200 1878
   Rk              Player Pos Age  Tm  G GS   MP  FG  FGA X3P X3PA X2P X2PA  FT FTA ORB DRB TRB AST STL BLK TOV  PF  PTS
20 19 Giannis Antetokounmpo  PG  21 MIL 80 79 2823 513 1013  28  109 485  904 296 409 113 499 612 345  94 113 208 258 1350
     Rk      Player Pos Age  Tm  G GS   MP  FG FGA X3P X3PA X2P X2PA FT FTA ORB DRB TRB AST STL BLK TOV  PF PTS
418 378 Rajon Rondo  PG  29 SAC 72 72 2537 355 782  62  170 293  612 87 150  77 358 435 839 141  10 278 175 859
     Rk        Player Pos Age  Tm  G GS   MP  FG  FGA X3P X3PA X2P X2PA  FT FTA ORB DRB TRB AST STL BLK TOV  PF  PTS
111 105 Stephen Curry  PG  27 GSW 79 79 2700 805 1598 402  886 403  712 363 400  68 362 430 527 169  15 262 161 2375
     Rk    Player Pos Age  Tm  G GS   MP  FG  FGA X3P X3PA X2P X2PA  FT FTA ORB DRB TRB AST STL BLK TOV  PF  PTS
496 446 John Wall  PG  25 WAS 77 77 2784 572 1349 115  328 457 1021 272 344  42 337 379 789 145  59 318 159 1531
```

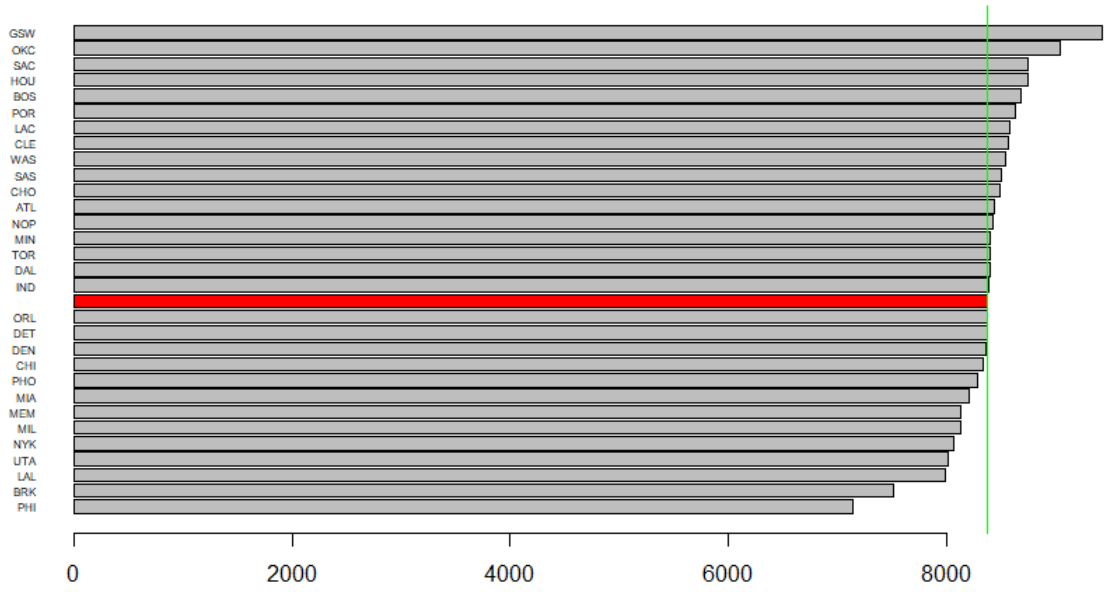
14. Από το ερώτημα 3, αποθηκεύσαμε στο διάνυσμα s το συνολικό αριθμό πόντων που σκόραρε η κάθε ομάδα. Εδώ, θα κατασκευάσουμε ένα νέο διάνυσμα, το οποίο θα περιέχει το διάνυσμα s μαζί με την μέση τιμή του s και θα το ονομάσουμε b.Έπειτα, θα κατασκευάσουμε ένα διάνυσμα a, στο οποίο θα συμπεριλάβουμε τα ονόματα τον ομάδων, μαζί με ένα ¨κενό¨, που θα αντιστοιχεί στη μέση τιμή. Τέλος, με τη βοήθεια αυτού των διανυσμάτων a, b, θα δημιουργήσουμε ένα data frame, με τη χρήση του οποίου θα φτιάξουμε το ζητούμενο barplot:

```
> a<-c(levels(nbadata$Tm),"")
> b<-c(s,mean(s))
> Data<-data.frame(a,b)
> library(plyr)
> Data<-data.frame(a,b)
> Data<-arrange(Data,b)
> Data
```

	Team	Points
1	PHI	7142
2	BRK	7503
3	LAL	7982
4	UTA	8010
5	NYK	8065
6	MIL	8122
7	MEM	8126
8	MIA	8204
9	PHO	8271
10	CHI	8335
11	DEN	8355
12	DET	8361
13	ORL	8369
14		8370
15	IND	8377
16	DAL	8388
17	TOR	8394
18	MIN	8398
19	NOP	8423
20	ATL	8433
21	CHO	8479
22	SAS	8490
23	WAS	8534
24	CLE	8554
25	LAC	8569
26	POR	8622
27	BOS	8669
28	HOU	8737
29	SAC	8740
30	OKC	9038
31	GSW	9421

Με τον πίνακα αυτόν, παρατηρούμε ότι ο μέσος όρος πόντων βρίσκεται στην 14 θέση.

```
> barplot(Data$b,horiz = TRUE,names.arg =
Data$a,cex.names=0.5,col=c(rep('gray',13),'red',rep('gray',17)),las=1)
> abline(v = mean(s), col = 'green')
```



15. Θα υπολογίσουμε τον αριθμό των παικτών που σκόραραν πάνω από 1000 πόντους με τον εξής αλγόριθμο:

```
> c<-0
> for(i in 1:dim(nbadata)[1]) {
+ if((nbadata[i,25])>1000) {
+ c<-c+1
+ }
+ }
> print(c)
```

Το αποτέλεσμα που λαμβάνουμε είναι ότι **76** παίκτες σκόραραν πάνω από 1000 πόντους. Όστοσο, στα στοιχεία που μας δίνονται, μπορεί να παρατηρήσει κάποιος ότι ο παίκτης Tobias Harris αναγράφεται 2 φορές, καθώς τη σεζόν αυτή έπαιξε με 2 διαφορετικές ομάδες. Συνολικά, ο παίκτης αυτός πέτυχε 1116 πόντους, επομένως οι παίκτες με συνολικό αριθμό πόντων άνω των 1000 ανέρχονται στους **77**. Από τους υπόλοιπους παίκτες, των οποίων τα στοιχεία εμφανίζονται τουλάχιστον 2 φορές, κανένας δεν έχει πετύχει συνολικά πάνω από 1000 πόντους(πραγματοποιώντας έλεγχο όπως στα προηγούμενα ερωτήματα, όπου αθροίζουμε τα στατιστικά των παικτών αυτών).

16. Αρχικά, θα προσθέσουμε τα στοιχεία των παικτών, οι οποίοι έπαιξαν σε διαφορετικές ομάδες μέσα στη σαιζόν, με τρόπο όμοιο όπως σε προηγούμενα ερωτήματα. Ας

ονομάσουμε το ενμερωμένο data frame = newdata2. Έπειτα, θα δημιουργήσουμε ένα νέο data frame, στο οποίο θα συμπεριλάβουμε μόνο τους PG, PF, ως εξής:

```
> vpf<-NULL
> vpg<-NULL {
> for(i in 1:dim(newdata2)[1]) {
+ newdata2$Pos[i]==levels(newdata2$Pos)[2]) { vpf¡-c(vpf,i) }
+ }
> for(j in 1:dim(newdata2)[1]) {
+ newdata2$Pos[j]==levels(newdata2$Pos)[3]) { vpg¡-c(vpg,j) }
+ }
> pgpfdata<-newdata2[c(vpg,vpf),]
```

Με την εντολή `> pgpfdata2<-arrange(pgpfdata,pgpfdata$Pos)` θα ανακατατάξουμε τα δεδομένα μας ώστε στις πρώτες θέσεις να βρίσκονται οι PF και έπειτα οι PG.
Στη συνέχεια, θα ορίσουμε τις σταθερές npf, npg, D, ppg οι οποίες θα αποτελούν των αριθμό των PF, τον αριθμό των PG, τη διάσταση του data frame και τον αριμό της σειράς από τον οποία ξεκινούν οι PG.
Έπειτα, θα ελέγξουμε τις προϋποθέσεις για να κάνουμε ένα t-test. Αρχικά, θα πρέπει να ελέγξουμε την κανονικότητα των δειγμάτων. Με τον έλεγχο Shapiro-Wilk έχουμε:

```
> shapiro.test(pgpfdata2$PTS[ppg:D])

        Shapiro-Wilk normality test

data:  pgpfdata2$PTS[ppg:D]
W = 0.89088, p-value = 4.847e-07

> shapiro.test(pgpfdata2$PTS[1:npf])

        Shapiro-Wilk normality test

data:  pgpfdata2$PTS[1:npf]
W = 0.91845, p-value = 1.062e-05
```

Και στις δύο περιπτώσεις προκύπτει ότι p-value<0.05, επομένως και τα δύο δείγματα δεν ακολουθούν την κανονική κατανομή. Ωστόσο, επειδή έχουμε πολύ μεγάλο δείγμα (n>30), ο επόμενως έλγχος που θα κάνουμε θα αφορά την ομοσκεδαστικότητα. Με Bartlett έχουμε:

```
> bartlett.test(PTS ~ Pos, data=pgpfdata2)

        Bartlett test of homogeneity of variances

data:  PTS by Pos
Bartlett's K-squared = 6.4841, df = 1, p-value = 0.01088
```

Βλέπουμε ότι p-value<0.05, επομένως η ομοσκεδαστικότητα απορρίπτεται. Συνεπώς, οδηγούμαστε σε μη-παραμετρικό έλεγχο Mann-Whitney U-Test, για να ελέγξουμε αν οι διάμεσοι είναι ίσοι.

```
> wilcox.test(PTS ~ Pos, data=pgpfdata2)
```

```
        Wilcoxon rank sum test with continuity correction

data:  PTS by Pos
W = 4347, p-value = 0.2569
alternative hypothesis: true location shift is not equal to 0
```

Συνεπώς, η μηδενική υπόθεση, που είναι ότι οι πόντοι των PG, PF ανήκουν στον ίδιο πληθυσμό, δεν απορρίπτεται σε επίπεδο σημαντικότητας 5% ($p > 0.05$), συνεπώς οι PG, PF σκοράρουν το ίδιο.
Ως προς τους μέσους όρους, αφού ναι μεν το Shapiro test έδειξε ότι ασχολούμαστε με μη-κανονικά δεδομένα, επειδή έχουμε μεγάλο δείγμα, μπορούμε να το θεωρήσουμε ότι τα δεδομένα μας κατανέμονται κανονικά. Έπειτα, ελέγχουμε τις διασπορές:
> var.test(pgpfdata2$PTS[ppg:D],pgpfdata2$PTS[1:npf])

```
        F test to compare two variances

data:  pgpfdata2$PTS[ppg:D] and pgpfdata2$PTS[1:npf]
F = 1.6801, num df = 100, denom df = 100, p-value = 0.01007
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 1.132715 2.492011
sample estimates:
ratio of variances
          1.680101
```

Παρατηρούμε ότι οι διασπορές δεν είναι ίσες, συνεπώς έχουμε:
> t.test(PTS Pos,data=pgpfdata2,var.eq=FALSE)

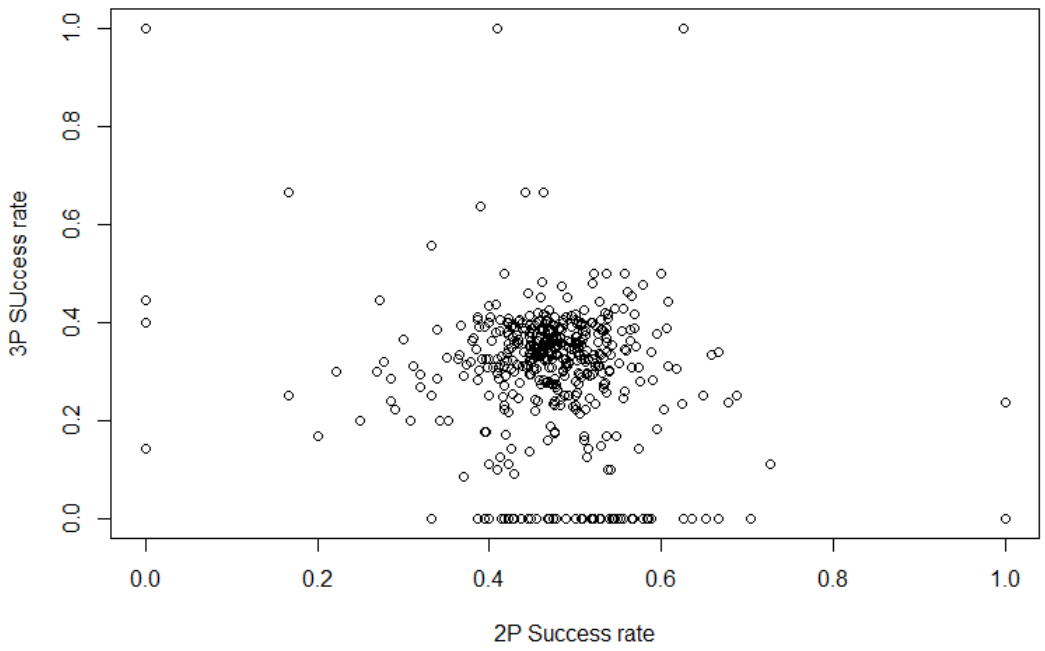
```
        Welch Two Sample t-test

data:  PTS by Pos
t = -1.592, df = 176.6, p-value = 0.1132
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -233.48033   24.98341
sample estimates:
mean in group PF mean in group PG
        474.2673         578.5158
```

Από τα παραπάνω προκύπτει ότι, σε επίπεδο σημαντικότητας 5%, το μέσο σκορ των PG δε διαφέρει σε βαθμό στατιστικά σημαντικό από το μέσο σκορ των PF ($p > 0.05$). Άρα, PG, PF σκοράρουν το ίδιο, κατά μέσο όρο.

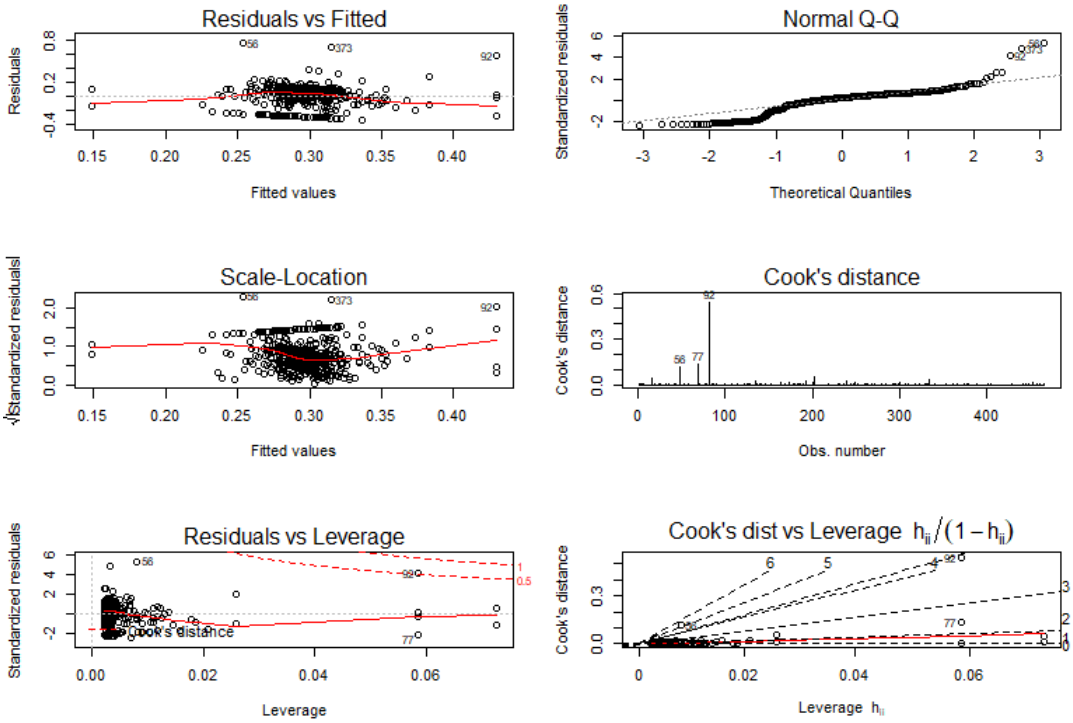
17. Ξεκινώντας, θα θέσουμε ως p_1, p_2 τα ποσοστά των σουτ 2 πόντων και 3 πόντων και θα θέσουμε ως model1 την παλινδόμιση μεταξύ των p_1, p_2.
> p1<-nbadata$X2P/nbadata$X2PA
> p2<-nbadata$X3P/nbadata$X3PA
> model1<-lm(p2 ∼ p1, data= nbadata)
Αρχικά θα κάνουμε έναν έλεγχο αν τα δεδομένα μας ακολουθούν την κανονική κατανομή:

```
        Shapiro-Wilk normality test

data:  p1
W = 0.83807, p-value < 2.2e-16

        Shapiro-Wilk normality test

data:  p2
W = 0.84989, p-value < 2.2e-16
```

Παρατηρούμε ότι κανενα από τα δύο δείγματα δεν ακολουθούν την κανονική κατανομή ($p < 0.05$). Ας δούμε επίσης κάποια ενδεικτικά γραφήμματα:



Από το γράφημα αυτό, είναι ενδεικτικό ότι δεν υπάρχει γραμμικότητα μεταξύ του Y και του X



Με εντολή summary(model1) θα δούμε μερικά στοιχεία για το μοντέλο μας:

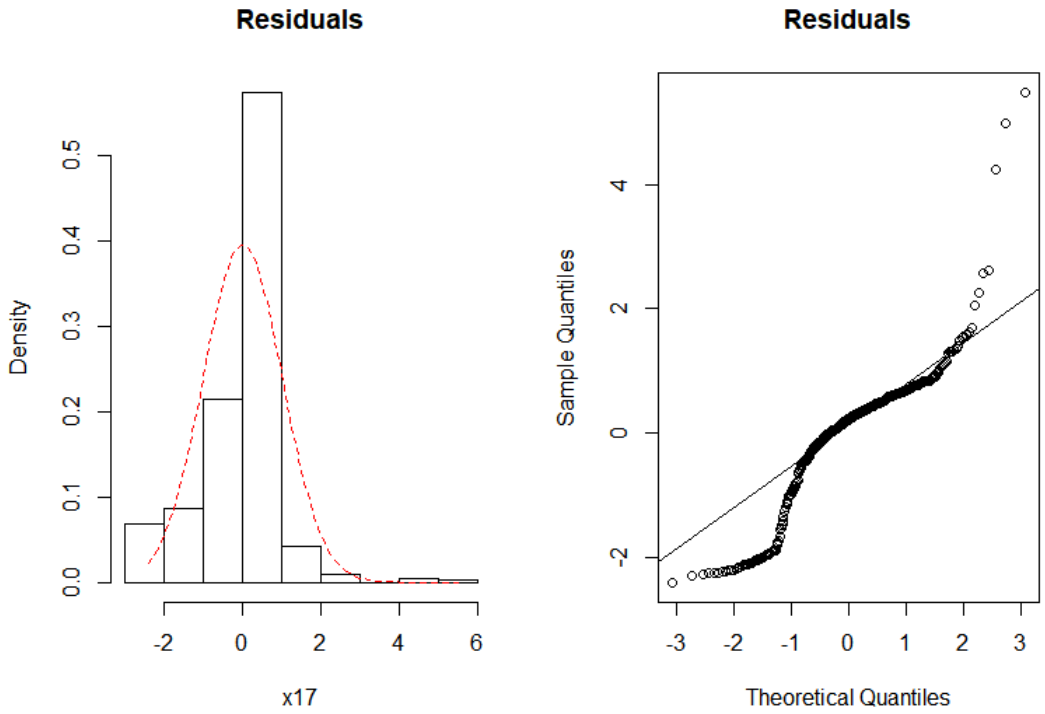
```
Call:
lm(formula = p2 ~ p1, data = nbadata)

Residuals:
     Min       1Q   Median       3Q      Max
-0.33622 -0.04529  0.03296  0.08050  0.74609

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.43000    0.03409  12.614  < 2e-16 ***
p1          -0.28134    0.07088  -3.969 8.34e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1409 on 465 degrees of freedom
  (58 observations deleted due to missingness)
Multiple R-squared:  0.03277,   Adjusted R-squared:  0.03069
F-statistic: 15.76 on 1 and 465 DF,  p-value: 8.344e-05
```

Από τη στήλη Estimate υπολογίζουμε τα $b_0 b_1$, παρατηρούμε ότι το μοντέλο μας θα έχει την μορφή:

$$Y = 0.43 - 0.28134X + \epsilon$$

όπου το U θα αντιστοιχει στο p_2, το X στο p_1 και $\epsilon \sim N(0, 0.48^2)$.
Έχοντας υπόθεση $H_0 : b_1 = 0 vs. H_1 : b_1 \neq 0$, για το b_1 βλέπουμε ότι $p < 0.05$ επομένως απορρίπτουμε τη μηδενική υπόθεση, επομένως το p_2 εξαρτάται από το p_1 και μάλιστα η

σχέση τους είναι αρνητική καθώς $b_1 < 0$. Ομοίως, για το b_0, έχουμε ότι $p < 0.05$ επομένως και εδώ τη μηδενική υπόθεση ότι $b_0 = 0$. Επιπ΄ροσθετα βλέπουμε ότι $R^2 < 0.7$ συνεπώς προβλέπουμε ότι δεν υπάρχει μεγάλη συσχέτιση μεταξύ των δύο μεταβλητών. Ξεκινάμε να κάνουμε έλεγχο, λοιπόν, για τα *residuals*.
Για την κανονικότητα έχουμε:
> shapiro.test(rstudent(model1))




```
        Shapiro-Wilk normality test

data:  rstudent(model1)
W = 0.87994, p-value < 2.2e-16
```

Είναι εμφανές ότι δεν υπάρχει κανονικότητα στα studentized residuals.
Για την ανεξαρτησία, θα χρησιμοποιήσουμε τον έλεγχο Durbin - Watson:
> dwtest(model1)

```
        Durbin-Watson test

data:  model1
DW = 2.0653, p-value = 0.7607
alternative hypothesis: true autocorrelation is greater than 0
```

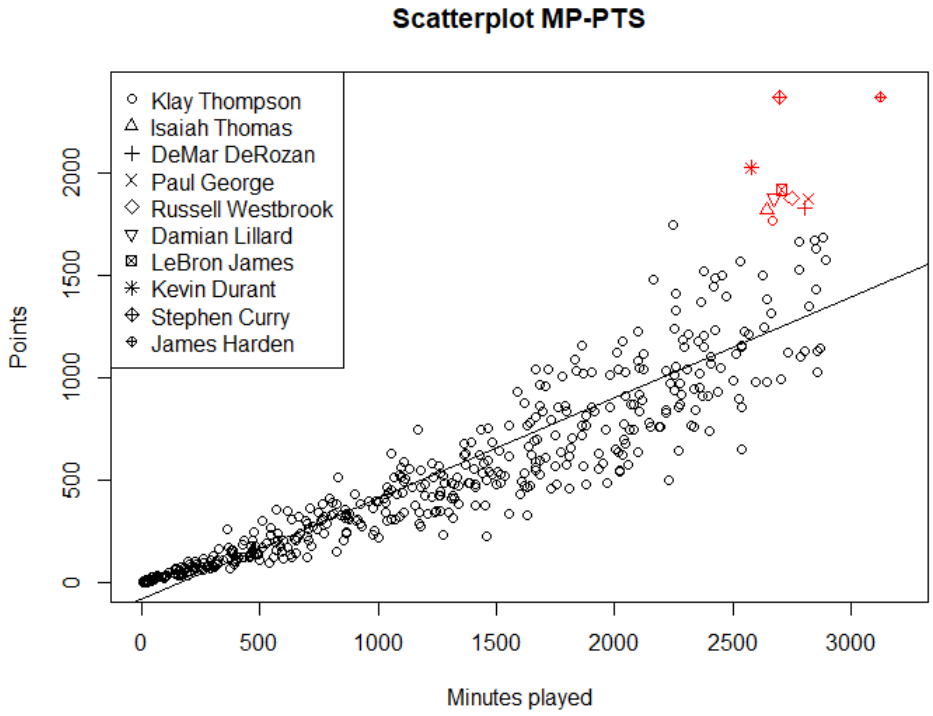
Βλέπουμε ότι ο δείκτης Durbin - Watson είναι ίσος με $2.0653 \simeq 2$ με τον σχετικό

έλεγχο να μην απορρίπτει τη μηδενική υπόθεση

$$H_0 : d = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=2}^{n} e_i} = 0$$

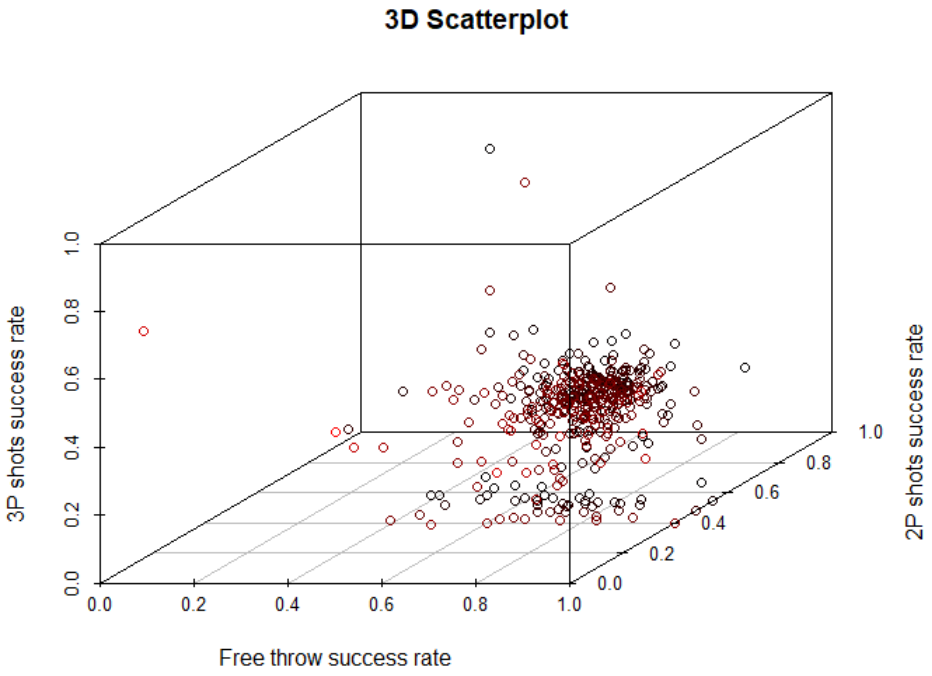
σε επίπεδο $a = 5\%$, αφού $p > 0.05$ και συνεπώς να μην εντοπίζεται κάποιο πρόβλημα προϋπόθεσης της ανεξαρτησίας των residuals

18. (α΄) Στο πρώτο γράφημα θα παραστήσουμε τη σχέση μεταξύ του χρόνου που έπαιξαν οι παίκτες με τους πόντους που πέτυχαν.

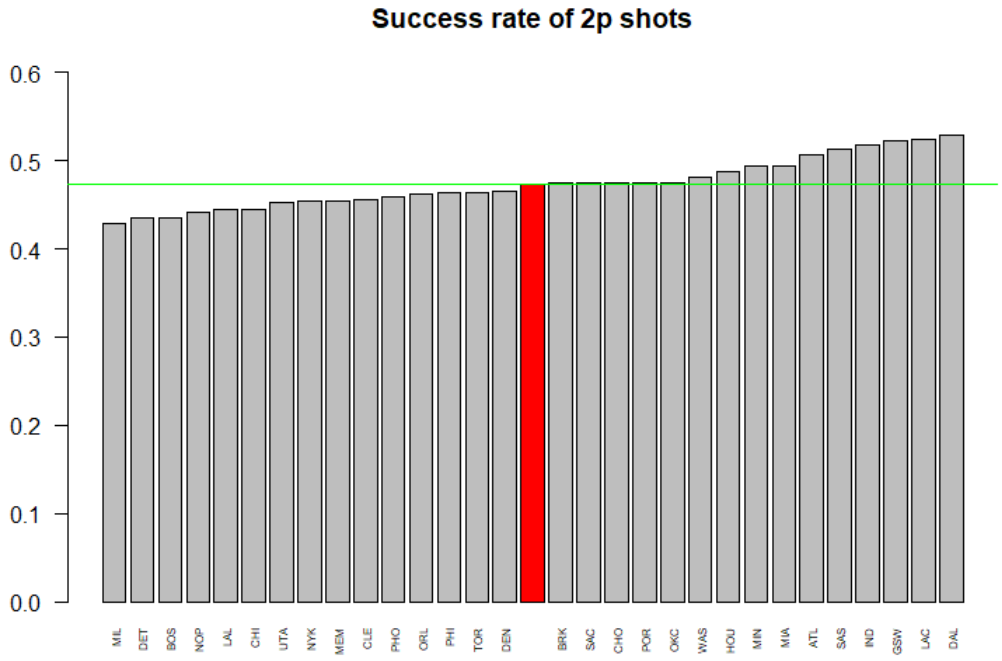


Όπως ήταν αναμενόμενο, οι παίκτες με το μεγαλύτερο χρόνο συμμετοχής είναι και οι πιο παραγωγικοί

(β΄) Στο επόμενο γράφημα θα δούμε τη σχέση της ευστοχίας των παικτών στις βολές, στα σουτ 2 πόντων καθώς και στα σουτ 3 πόντων.

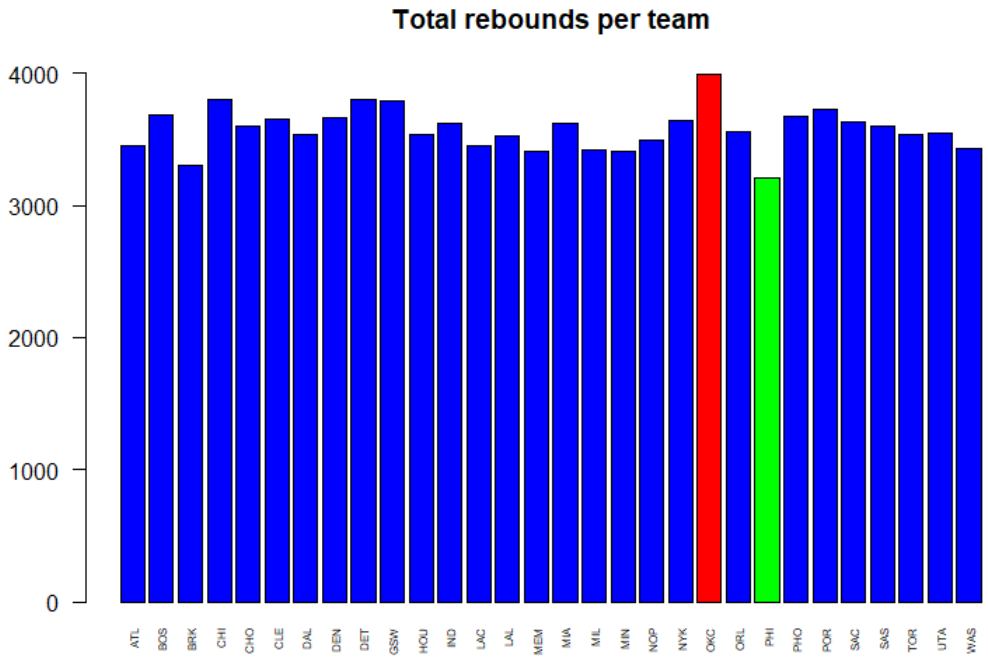


(γ΄) Στο γράφημα που ακολουθεί θα παρουσιάσουμε την ευστοχία των ομάδων στα σουτ των 2 πόντων.

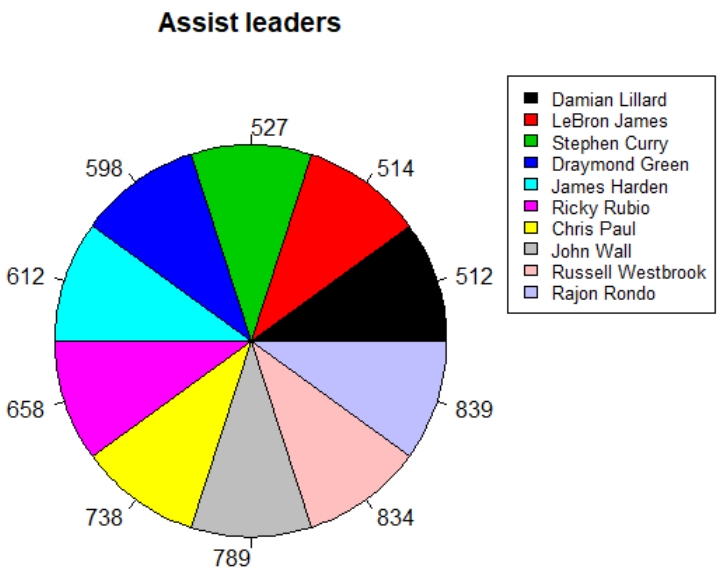


Ενδιαφέρον προκαλεί το γεγονός ότι η πρωταθλήτρια ομάδα για το 2016 (Cleveland Cavaliers) είχαν ποσοστό ευστοχίας μικρότερο από το μέσο όρο.

(δ΄) Στο τέταρτο γράφημα θα παρουσιάσουμε τα rebounds που μάζεψε συνολικά κάθε ομάδα.



(ε΄) Στο τελευταίο γράφημα θα παρουσιάσουμε τους 10 παίκτες με τις περισσότερες assists.



Ενδιαφέρον προκαλεί το γεγονός ότι από τους 10 αυτούς παίκτες, οι 6 είναι PG, 2 είναι SG μαζί με έναν SF και έναν PF, και μάλιστα 2 παίκτες ανήκουν στην ίδια ομάδα (Stephen Curry & Draymond Green).