ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

ΣΧΟΛΗ ΕΠΙΣΤΗΜΩΝ & ΤΕΧΝΟΛΟΓΙΑΣ ΤΗΣ ΠΛΗΡΟΦΟΡΙΑΣ
SCHOOL OF INFORMATION SCIENCES & TECHNOLOGY
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ
DEPARTMENT OF INFORMATICS

ΣΤΑΤΙΣΤΙΚΗ ΣΤΗΝ ΠΛΗΡΟΦΟΡΙΚΗ

Εξερευνητική Ανάλυση Δεδομένων

Εργασία 1

Μαρία Σχοινάκη

Σοφία Παπαϊωάννου

Άσκηση 1

a) Υπολογίζουμε τη **σύνοψη των 5 αριθμών** (*ελάχιστο, πρώτο τεταρτημόριο Q1, διάμεσος, τρίτο τεταρτημόριο Q3, μέγιστο*), τη **μέση τιμή** και την **τυπική απόκλιση** για κάθε ομάδα δεδομένων, έτσι ώστε να χρησιμοποιηθούν στα επόμενα ερωτήματα.

Δεδομένα I

Σύνοψη 5 αριθμών = (min=**30.3**, max=**34.5**, med=**32.65**, Q1=**31.1**, Q3= **33.6**)

mean = $\frac{30.3 + 31.0 + 31.1 + 32.1 + 32.6 + 32.7 + 33.4 + 33.6 + 34.2 + 34.5}{10}$ = **32.55**

Τυπική Απόκλιση = 1.419898

Δεδομένα II

Σύνοψη 5 αριθμών = (min=**0.0**, max=**9.0**, med=**1.3**, Q1=**0.2**, Q3= **4.2**)

mean = $\frac{0.0 + 0.0 + 0.2 + 0.8 + 1.2 + 1.4 + 3.2 + 4.2 + 6.4 + 9.0}{10}$ = **2.64**

Τυπική Απόκλιση = 3.059121

Δεδομένα III

Σύνοψη 5 αριθμών = (min=**0.0**, max=**96.0**, med=**39.5**, Q1=**17.5**, Q3= **59.0**)

mean = $\frac{0 + 1 + 6 + 8 + 10 + \ldots + 87 + 88 + 89 + 94 + 96}{40}$ = **41.15**

Τυπική Απόκλιση = 28.26754

Stemplots

Δεδομένα I

30	3
31	0 1
32	1 6 7
33	4 6
34	2 5

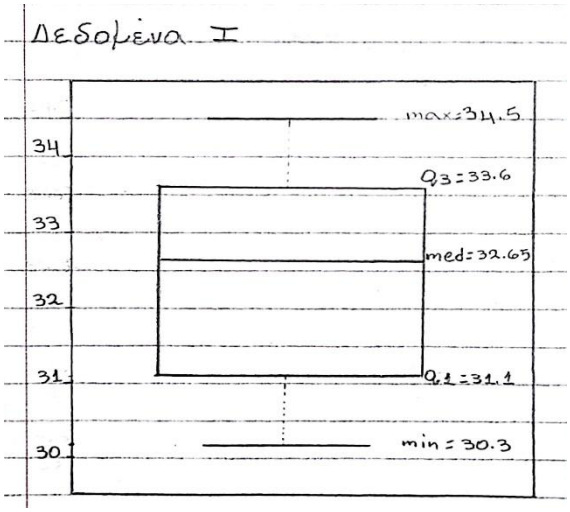
Δεδομένα II

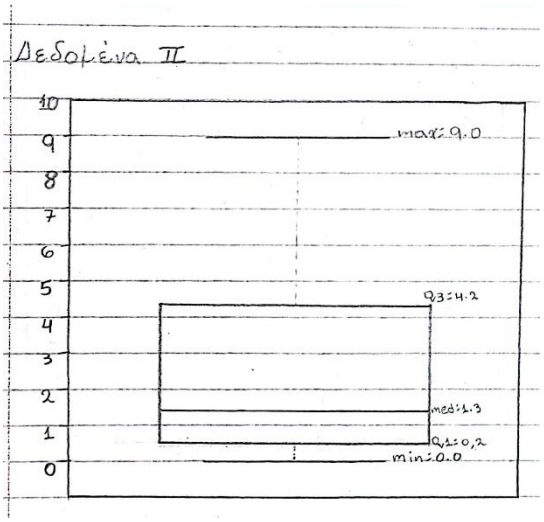
0	0 0 2 8
1	2 4
2	
3	2
4	2
5	
6	4
7	
8	
9	0

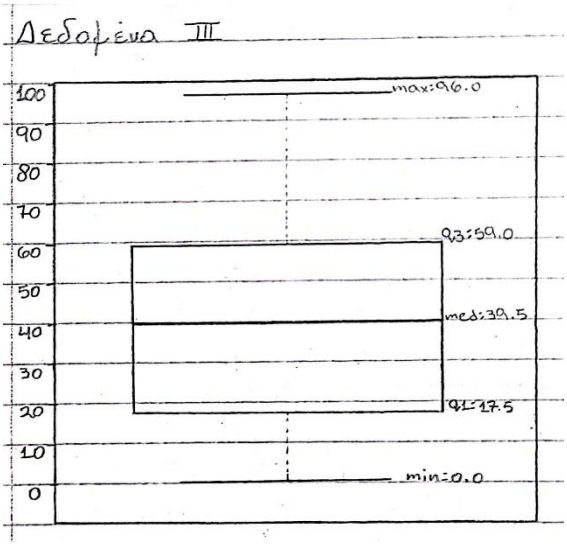
Δεδομένα III

0	0 1 6 8
1	0 3 5 6 7 7 8 8
2	0 0 1 5 6
3	0 5 9
4	0 1 3 4 6 8
5	2 4 8 9 9
6	0 6
7	
8	1 6 7 8 9
9	4 6

Boxplots







b) Για την **ομάδα δεδομένων Ι**, όπου τα δεδομένα είναι συμμετρικά κατανεμημένα γύρω από τη μέση τιμή $\bar{x}$ και δεν υπάρχουν ακραίες τιμές (*outliers*), η **μέση τιμή** και η **τυπική απόκλιση** είναι οι κατάλληλοι δείκτες για τη σύνοψη της κατανομής. Αυτοί οι στατιστικοί δείκτες περιγράφουν με ακρίβεια το κέντρο και τη διασπορά των δεδομένων.

Αντίθετα, οι **ομάδες δεδομένων ΙΙ και ΙΙΙ** έχουν μη συμμετρικές κατανομές και περιέχουν ακραίες τιμές. Αυτό καθιστά τη χρήση της μέσης τιμής και της τυπικής απόκλισης λιγότερο αποτελεσματική για την περιγραφή τους, καθώς οι δείκτες αυτοί επηρεάζονται σημαντικά από τις ακραίες τιμές.

Σε αυτές τις περιπτώσεις, η **σύνοψη των 5 αριθμών** αποτελεί καταλληλότερη μέθοδο περιγραφής. Η σύνοψη αυτή είναι ανθεκτική στις ακραίες τιμές και δίνει μια πιο αξιόπιστη εικόνα για τη θέση και την κατανομή των δεδομένων στις **ομάδες ΙΙ και ΙΙΙ**.

c) Για να αξιολογήσουμε αν τα δεδομένα ακολουθούν κανονική κατανομή, θα χρησιμοποιήσουμε τον κανόνα **68-95-99.7**, σύμφωνα με τον οποίο για κάθε κανονική κατανομή $N(\mu, \sigma)$, όπου μ = μέση τιμή = διάμεση τιμή και σ = τυπική απόκλιση ισχύει ότι :

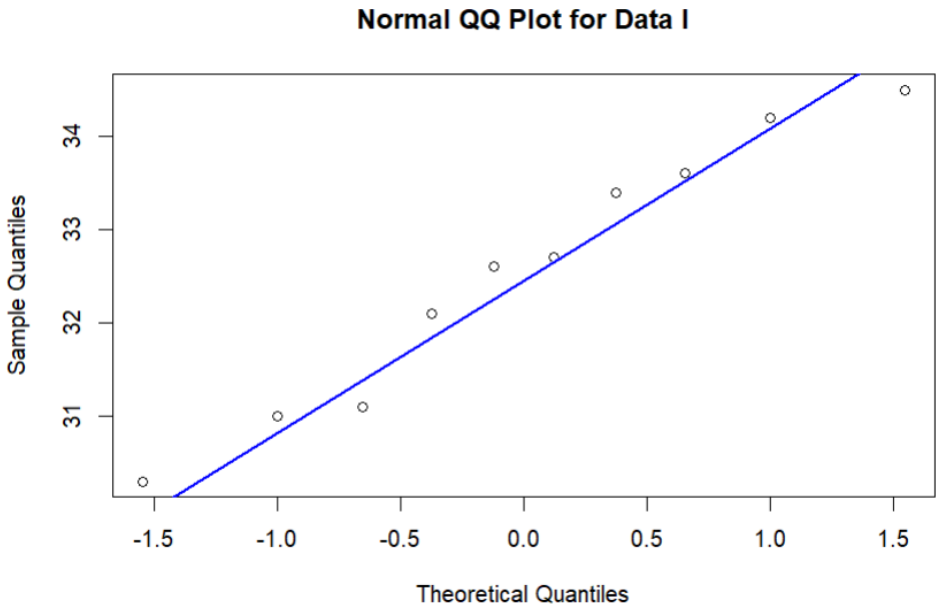
- το 68% των περιπτώσεων βρίσκεται στο διάστημα $(\mu - \sigma, \mu + \sigma)$
- το 95% των περιπτώσεων βρίσκεται στο διάστημα $(\mu - 2\sigma, \mu + 2\sigma)$
- Το 99.7% των περιπτώσεων βρίσκεται στο διάστημα $(\mu - 3\sigma, \mu + 3\sigma)$

Ένα άλλο εργαλείο που μπορούμε να χρησιμοποιήσουμε είναι το **Normal-quantile plot** (*ή QQ plot*). Πρόκειται για ένα οπτικό διάγραμμα που συγκρίνει την κατανομή των δεδομένων μας με την κανονική κατανομή. Στο QQ plot, τα δεδομένα μας τοποθετούνται στον άξονα y και τα θεωρητικά ποσοστημόρια της κανονικής κατανομής στον άξονα x.

- **Αν τα σημεία στο QQ plot ευθυγραμμίζονται** κοντά σε μια ευθεία γραμμή, τότε η κατανομή των δεδομένων μας ταιριάζει καλά με την κανονική κατανομή.
- **Αν τα σημεία αποκλίνουν σημαντικά από την ευθεία**, τότε η κατανομή των δεδομένων είναι πιθανώς διαφορετική από την κανονική.

<u>Δεδομένα Ι</u>

- **Στο διάστημα (μ - σ, μ + σ)**, δηλαδή στο (31.1301, 33.9699), βρίσκονται 5 από τις 10 τιμές, δηλαδή το **50%** των δεδομένων. Κανονικά, θα έπρεπε να είχαμε **68%** των δεδομένων, επομένως έχουμε απόκλιση **-18%**.
- **Στο διάστημα (μ - 2σ, μ + 2σ)**, δηλαδή στο (29.7102, 35.3898), βρίσκονται και οι 10 από τις 10 τιμές, δηλαδή το **100%** των δεδομένων. Κανονικά, θα έπρεπε να είχαμε **95%**, οπότε υπάρχει απόκλιση **+5%**.
- **Στο διάστημα (μ - 3σ, μ + 3σ)**, δηλαδή στο (28.29031, 36.80969), βρίσκονται και πάλι και οι 10 από τις 10 τιμές, δηλαδή το **100%** των δεδομένων, ενώ κανονικά θα έπρεπε να είχαμε **99.7%**. Άρα η απόκλιση είναι μόνο **+0.03%**.

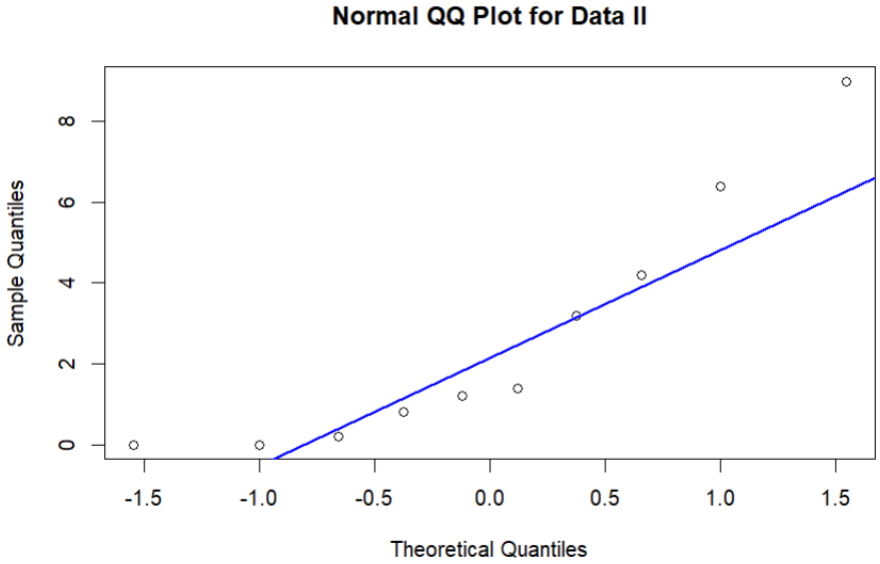


Παρατηρούμε ότι τα σημεία στο **QQ plot της Ομάδας Δεδομένων I** είναι αρκετά ευθυγραμμισμένα, δηλαδή βρίσκονται κοντά στην ευθεία αναφοράς. Αυτό δείχνει ότι τα δεδομένα της ομάδας αυτής έχουν μια κατανομή που μοιάζει με την κανονική, καθώς η ευθυγράμμιση των σημείων υποδηλώνει συμμετρία και ομαλότητα στην κατανομή τους.

Επιπλέον, οι αποκλίσεις από τον κανόνα **68-95-99.7** είναι μικρές, πράγμα που ενισχύει την παρατήρησή μας. Με βάση την ευθυγράμμιση των σημείων και τα σχετικά χαμηλά ποσοστά αποκλίσεων, μπορούμε να συμπεράνουμε ότι η κατανομή των δεδομένων της **Ομάδας I** προσεγγίζει αρκετά καλά την κανονική κατανομή.

Δεδομένα II

- **Στο διάστημα (μ- σ, μ + σ)** δηλαδή το (-0.41912, 5.6991) ανήκουν οι 8 από τις 10 τιμές της ομάδας, δηλαδή το **80%**. Σύμφωνα με την κανονική κατανομή θα έπρεπε να είχαμε το 68%, οπότε έχουμε απόκλιση **+12%**.
- **Στο διάστημα (μ- 2σ, μ + 2σ)** δηλαδή το (-3.478242, 8.758242) ανήκουν οι 9 από τις 10 τιμές της ομάδας, δηλαδή το **90%**. Σύμφωνα με την κανονική κατανομή θα έπρεπε να είχαμε το **95%**, οπότε έχουμε απόκλιση **-5%**.
- **Στο διάστημα (μ- 3σ, μ + 3σ)** δηλαδή το (-6.537363, 11.81736) ανήκουν και οι 10 από τις 10 τιμές της ομάδας, δηλαδή το **100%**. Σύμφωνα με την κανονική κατανομή θα έπρεπε να είχαμε το **99,7%**, οπότε έχουμε απόκλιση **+0,03%**

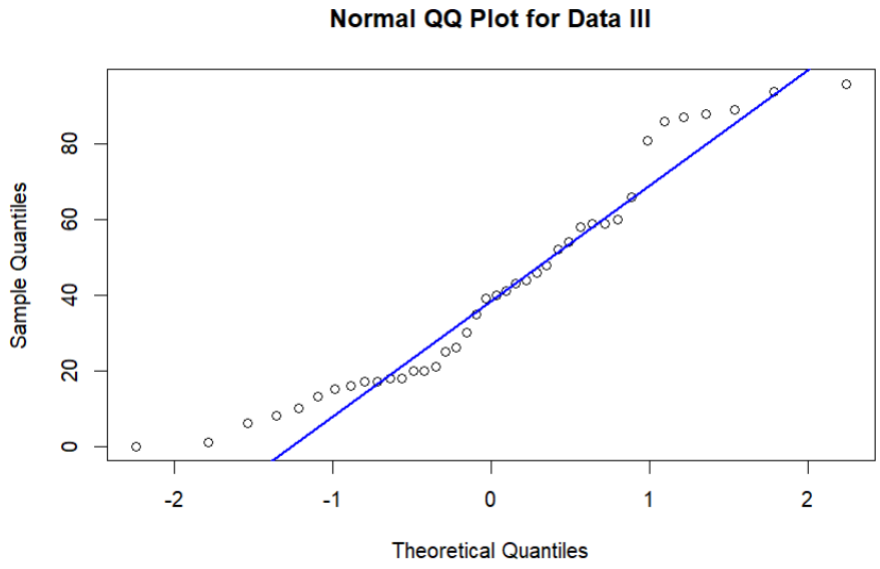


Παρατηρούμε ότι τα σημεία στο **QQ plot της Ομάδας Δεδομένων II** δεν ευθυγραμμίζονται απόλυτα με την ευθεία αναφοράς, καθώς παρουσιάζουν μια καμπύλη διάταξη που θυμίζει κυρτή συνάρτηση. Αυτό δείχνει ότι τα δεδομένα της **Ομάδας II** έχουν κάποια απόκλιση από την κανονική κατανομή, καθώς η μη ευθυγράμμιση των σημείων υποδηλώνει ασυμμετρία.

Συνεπώς, από το διάγραμμα μπορούμε να συμπεράνουμε ότι η κατανομή των δεδομένων της **Ομάδας II** δεν ταιριάζει τόσο καλά με την καμπύλη πυκνότητας της κανονικής κατανομής.

Δεδομένα III

- **Στο διάστημα (μ- σ, μ + σ)** δηλαδή το (12.88246, 69.41754) ανήκουν οι 28 από τις 40 τιμές της ομάδας, δηλαδή το **70%**. Σύμφωνα με την κανονική κατανομή θα έπρεπε να έχουμε το **68%**, οπότε υπάρχει απόκλιση **+2%**.
- **Στο διάστημα (μ- 2σ, μ + 2σ)** δηλαδή το (-15.38508, 97.68508) ανήκουν και οι 40 από τις 40 τιμές της ομάδας, δηλαδή το **100%**. Σύμφωνα με την κανονική κατανομή θα έπρεπε να έχουμε το **95%**, οπότε έχουμε απόκλιση **+5%**.
- **Στο διάστημα (μ- 3σ, μ + 3σ)** δηλαδή το (-43.65262, 125.9526) ανήκουν και οι 40 από τις 40 τιμές της ομάδας, δηλαδή το **100%**. Σύμφωνα με την κανονική κατανομή θα έπρεπε να είχαμε το **99.7%**, οπότε έχουμε απόκλιση **+0,03%**.



Παρατηρούμε ότι στο **QQ plot της Ομάδας Δεδομένων III**, περίπου το ένα τρίτο των σημείων, κυρίως τα κεντρικά σημεία, ευθυγραμμίζονται με την ευθεία αναφοράς. Αυτή η μερική ευθυγράμμιση υποδηλώνει ότι μόνο το κεντρικό τμήμα της κατανομής πλησιάζει την κανονική κατανομή, ενώ τα υπόλοιπα σημεία αποκλίνουν σημαντικά.

Η έλλειψη συνολικής ευθυγράμμισης δείχνει ότι τα δεδομένα της **Ομάδας III** δεν ακολουθούν πλήρως την καμπύλη πυκνότητας της κανονικής κατανομής. Συνολικά, μπορούμε να συμπεράνουμε ότι η κατανομή της **Ομάδας III** δεν προσεγγίζει ικανοποιητικά την κανονική κατανομή.

Άσκηση 2

a) Τα δεδομένα που χρησιμοποιούμε προέρχονται από το "Sports Reference | Sports Stats" και αφορούν τα στατιστικά των ομάδων του NBA για την κανονική περίοδο από τη σεζόν 1946-1947 έως και την πιο πρόσφατη ολοκληρωμένη σεζόν, (2023-24). Συνολικά, περιέχονται **1663 περιπτώσεις** (*γραμμές δεδομένων*). Σε ορισμένες περιπτώσεις, ενδέχεται να λείπουν δεδομένα, τα οποία καταγράφονται με την ένδειξη **NA**.

b) Κατηγορηματικές Μεταβλητές

Το dataset περιλαμβάνει **2** βασικές κατηγορηματικές μεταβλητές:

1. **Season**: Η σεζόν στην οποία αναφέρονται τα δεδομένα. Η μορφή της είναι **'x-x+1'**, όπου το "**x**" είναι το έτος και στο x+1 κρατάμε τα 2 τελευταία ψηφία. Για παράδειγμα, η σεζόν 2020 θα καταγραφεί ως '2020-21'.

2. **State**: Ο γεωγραφικός διαχωρισμός της ομάδας, με τις δύο πιθανές τιμές: **East** (Ανατολή) και **West** (Δύση). Αυτή η κατηγορία δείχνει σε ποιο από τα δύο βασικά περιφερειακά διαμερίσματα του NBA ανήκει η ομάδα.

Ποσοτικές Μεταβλητές

Το dataset περιλαμβάνει **25** ποσοτικές μεταβλητές, οι οποίες αφορούν διάφορα στατιστικά στοιχεία του NBA. Αυτές που θα μας απασχολήσουν είναι:

1. **W/L%** (*Ποσοστό Νικών/Ηττών*): Το ποσοστό των νικών σε σχέση με τις συνολικές προσπάθειες της ομάδας για τη σεζόν, το οποίο υπολογίζεται ως το ποσοστό νικών (*Wins*) προς τον συνολικό αριθμό αγώνων (*Games*).

2. **G** (*Συνολικοί Αγώνες*): Ο συνολικός αριθμός των αγώνων που έπαιξε η ομάδα κατά τη διάρκεια της συγκεκριμένης σεζόν.

3. **PTS** (*Συνολικοί Πόντοι*): Ο συνολικός αριθμός των πόντων που πέτυχε η ομάδα στη διάρκεια της σεζόν.

Επιπλέον, υπάρχουν και άλλες ποσοτικές μεταβλητές που αναφέρονται σε διάφορα στατιστικά του μπάσκετ.

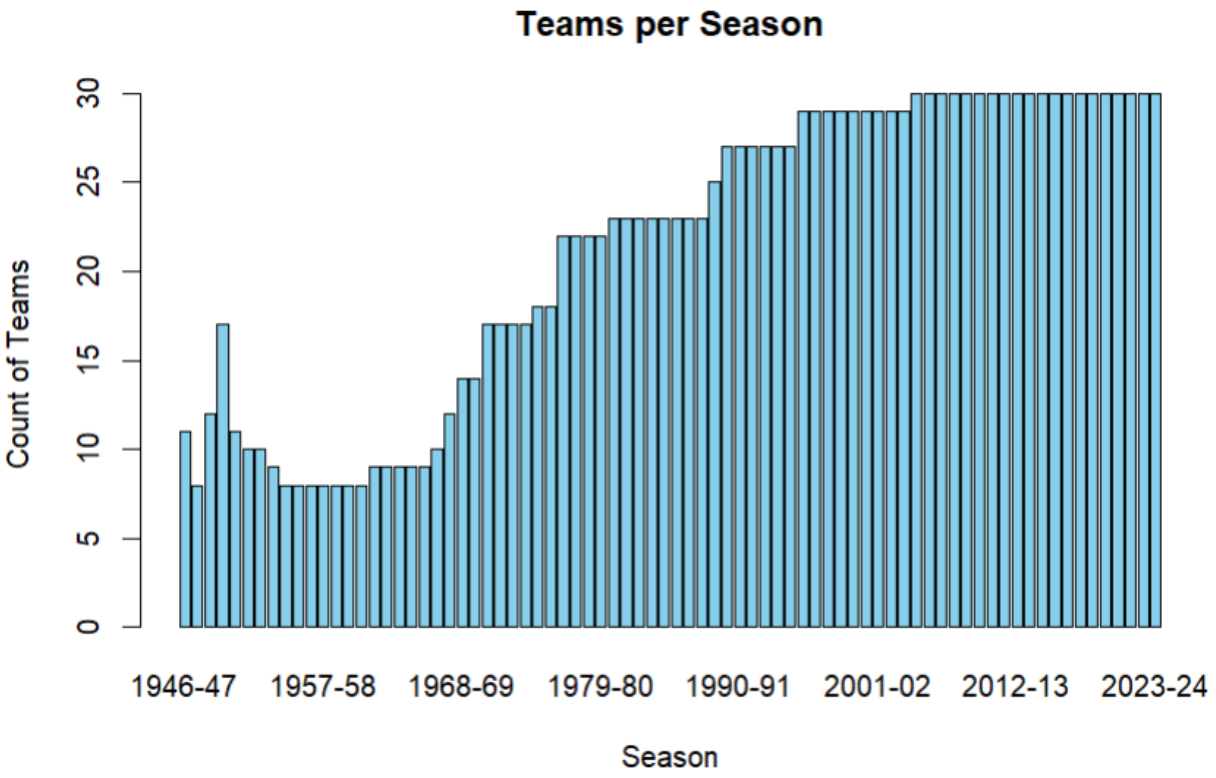
1. **FG** (*Συνολικά Εύστοχα Καλάθια*): Ο συνολικός αριθμός των εύστοχων προσπαθειών για καλάθια που πραγματοποίησαν οι παίκτες της ομάδας σε όλους τους αγώνες της σεζόν.

2. **MP** (*Συνολικά Λεπτά Παίξεων*): Ο συνολικός αριθμός λεπτών που αγωνίστηκαν όλοι οι παίκτες της ομάδας κατά τη διάρκεια της σεζόν. Αυτή η μεταβλητή δείχνει την ένταση της συμμετοχής της ομάδας στους αγώνες κ.α.

c) **Season**

Εντολές σε R

```
> barplot(table(season_vector),
+ main = 'Teams per Season',
+ xlab = 'Season',
+ ylab = 'Count of Teams',
+ col = 'skyblue')
```

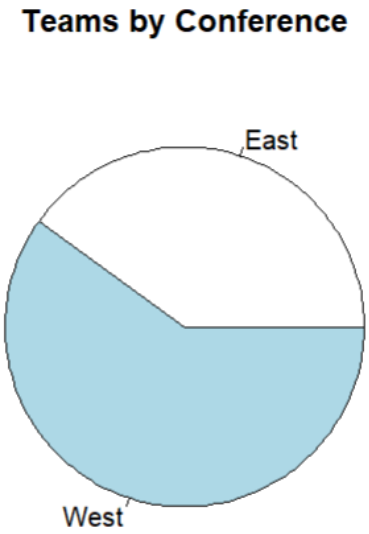


Η παραπάνω γραφική παράσταση (*barplot*) απεικονίζει την κατανομή της κατηγορηματικής μεταβλητής "**Season**", η οποία αναφέρεται στον αριθμό των ομάδων του NBA σε κάθε σεζόν από το 1946 μέχρι το 2023. Στην αρχή του πρωταθλήματος, το NBA ξεκίνησε με μόλις 11 ομάδες, και με την πάροδο των χρόνων προστίθενται συνεχώς νέες ομάδες, φτάνοντας τελικά στις 30 ομάδες, αριθμός που παραμένει σταθερός στις τελευταίες δύο δεκαετίες. Ενδιαφέρον έχει το γεγονός ότι στη σεζόν 1949-50 υπήρξαν 17 ομάδες, αριθμός που ήταν σημαντικά υψηλότερος από αυτόν των γειτονικών σεζόν. Αυτό οφείλεται στην ένωση δύο ξεχωριστών κατηγοριών του μπάσκετ (National Basketball League - NBL και Basketball Association of America - BAA) σε μία, με σκοπό την ενίσχυση του πρωταθλήματος.

State

Εντολές σε R

```
> pie(table(state),
+ main = "Teams by Conference")
```

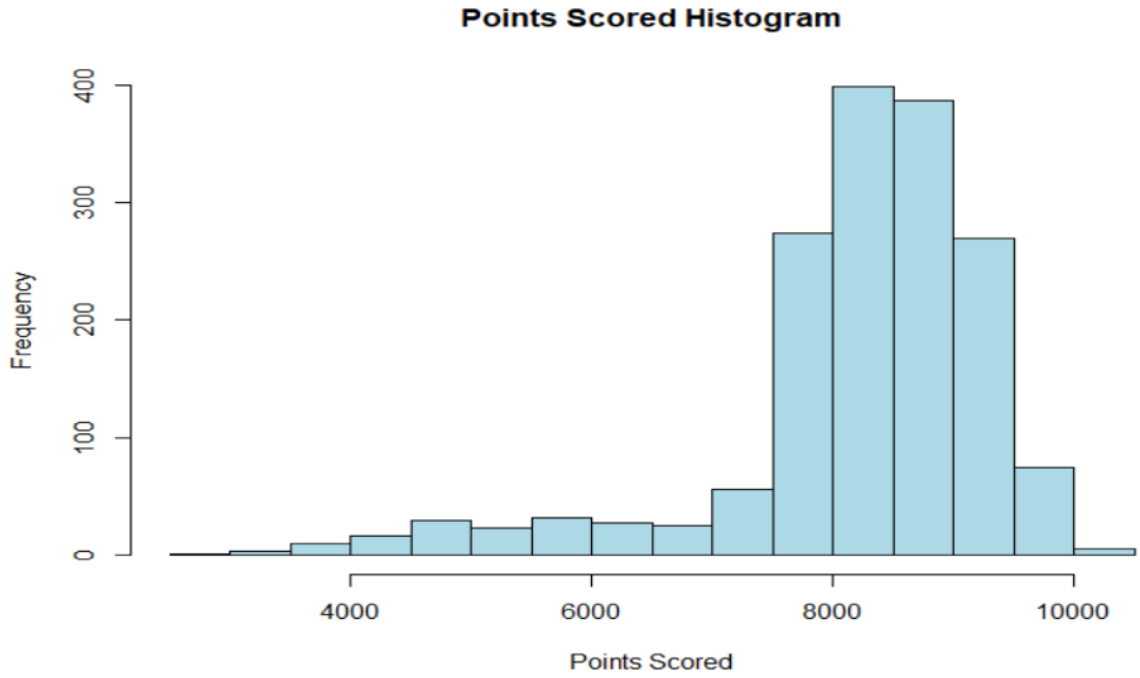


Στο παραπάνω διάγραμμα πίτας (pie chart), παρατηρούμε ότι σχεδόν το **60%** των ομάδων ανήκει στον όμιλο της Δύσης, ενώ το υπόλοιπο **40%** ανήκει στον όμιλο της Ανατολής. Η κατανομή αυτή αντανακλά τη γεωγραφική και ιστορική κατανομή των ομάδων στο NBA, με τη Δύση να έχει λίγο μεγαλύτερη εκπροσώπηση σε σχέση με την Ανατολή. Ο όμιλος της Δύσης έχει παραδοσιακά φιλοξενήσει περισσότερες ομάδες, ενώ η Ανατολή παραμένει με λιγότερες, αλλά εξίσου ανταγωνιστικές ομάδες.

Points Scored

Εντολές σε R

```
> hist(PTS,
+ main = "Points Scored Histogram",
+ xlab = "Points Scored")
```



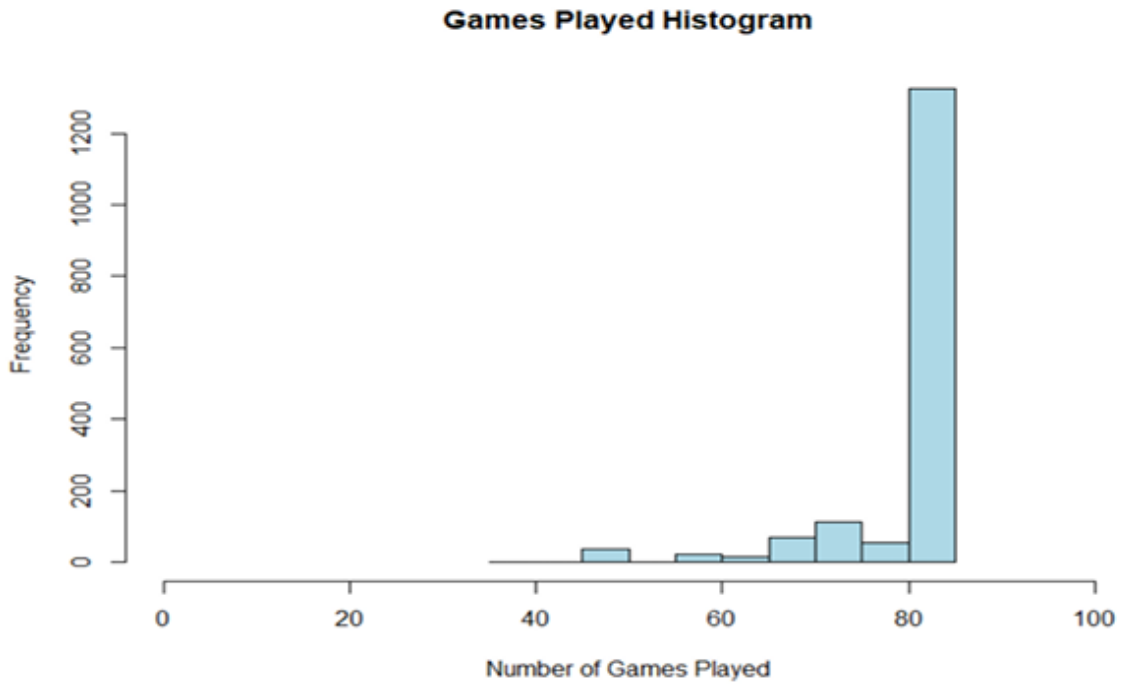
Παρατηρούμε ότι οι τιμές είναι συμμετρικά κατανεμημένες στο διάστημα Δ = (7000-10.000], υποδεικνύοντας μια κλασική κατανομή γύρω από μια κεντρική τιμή. Μια ενδιαφέρουσα παρατήρηση είναι ότι τα outliers, ή οι τιμές που απέχουν σημαντικά από τον κύριο όγκο των δεδομένων, πιθανότατα προκλήθηκαν από ομάδες που αποσύρθηκαν από το πρωτάθλημα. Αυτό είναι λογικό, καθώς λιγότεροι αγώνες συνήθως συνεπάγονται και λιγότερους συνολικούς πόντους, οδηγώντας σε τιμές εκτός του συνήθους εύρους.

Η μεγάλη συγκέντρωση των δεδομένων στο διάστημα Δ είναι αναμενόμενη, καθώς κάθε αγώνας NBA συνήθως παράγει περίπου 100 πόντους για κάθε ομάδα, με το σύνολο των αγώνων για κάθε ομάδα να ανέρχεται σε 82 παιχνίδια τη σεζόν. Επομένως, ένας μέσος όρος γύρω από 8000 πόντους συνολικά για την κάθε ομάδα είναι μια λογική τιμή.

Games

Εντολές σε R

```
> hist(G,
+ main = "Games Played Histogram",
+ xlab = "Number of Games Played",
+ xlim = c(0, 100))
```



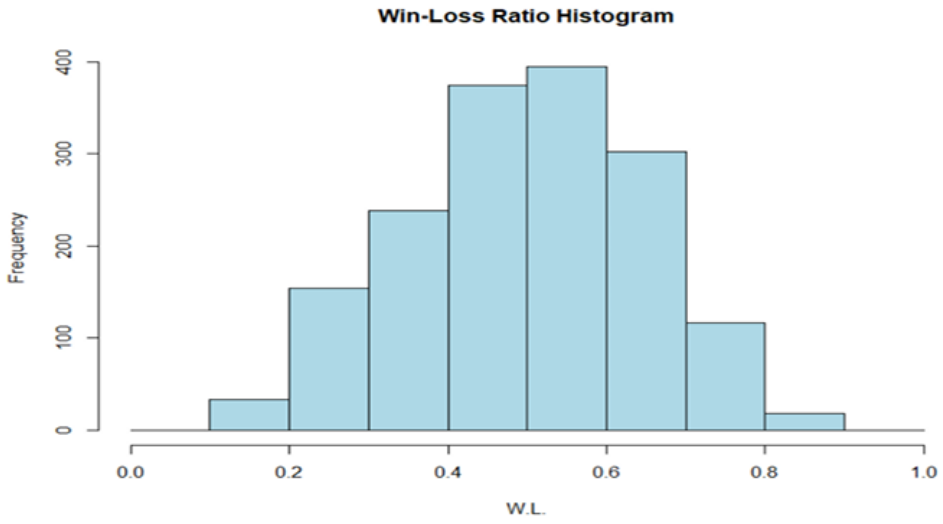
Από το παραπάνω ιστόγραμμα, παρατηρούμε μια πολύ μεγάλη συγκέντρωση συχνοτήτων στην τιμή 82, γεγονός που είναι αναμενόμενο, καθώς η κανονική διάρκεια μιας σεζόν στο NBA περιλαμβάνει 82 αγώνες για κάθε ομάδα. Αυτή η τιμή αντιπροσωπεύει τον πλήρη αριθμό αγώνων που παίζει μια ομάδα σε μια τυπική σεζόν.

Ωστόσο, παρατηρούμε και ορισμένα outliers, δηλαδή τιμές που απέχουν σημαντικά από τον κύριο όγκο των δεδομένων. Αυτά τα outliers πιθανότατα προκύπτουν από περιπτώσεις όπου κάποιες ομάδες απέσυραν την συμμετοχή τους από το πρωτάθλημα ή αντιμετώπισαν άλλες δυσκολίες που τις οδήγησαν να παίξουν λιγότερους αγώνες.

Win-Loss Ratio (W-L%)

Εντολές σε R

```
> hist(W.L.,
+ main = "Win-Loss Ration Histogram",
+ xlim = c(0, 1),
+ breaks = seq(0, 1, by = 0.1))
```



Από το παραπάνω ιστόγραμμα παρατηρούμε ότι η κατανομή των δεδομένων πλησιάζει πολύ την κανονική κατανομή. Αυτό εξηγείται από το γεγονός ότι, ενώ υπάρχουν λίγες ομάδες που έχουν εξαιρετικές επιδόσεις σε κάθε σεζόν με πολύ υψηλά ή πολύ χαμηλά ποσοστά νίκης-ήττας (πάνω από 80% ή κάτω από 20%), η πλειονότητα των ομάδων τείνει να κυμαίνεται σε ποσοστά νίκης-ήττας γύρω από το 40% με 60%. Για το λόγο αυτό, οι περισσότερες ομάδες συγκεντρώνονται γύρω από αυτές τις μέσες τιμές, δημιουργώντας έτσι τη μεγάλη συχνότητα στις κεντρικές περιοχές του διαγράμματος.

Υπάρχουν και άλλες πολλές μεταβλητές στο dataset αυτό, αλλά για τους σκοπούς της εργασίας περιοριστήκαμε στο να αναλύσουμε 2 από τις κατηγορηματικές και 3 από τις ποσοτικές.

d) Θα εξετάσουμε τις ποσοτικές μεταβλητές που έχουμε δώσει την μορφή των κατανομών τους στο προηγούμενο ερώτημα.

1. Points Scored (PTS)

Μέση Τιμή και Τυπική Απόκλιση: Η μέση τιμή (mean) των συνολικών πόντων υπολογίζεται στους 8398.49 πόντους, με τυπική απόκλιση 1153.32. Αυτές οι τιμές δίνουν μια γενική αίσθηση του μέσου όρου και της διακύμανσης των συνολικών πόντων. Δηλώνουν επίσης ότι οι περισσότερες ομάδες σημειώνουν κατά μέσο όρο περίπου 8000 πόντους σε μια κανονική σεζόν, με τυπική απόκλιση γύρω στους 1200 πόντους, μια διακύμανση που συμφωνεί και με την εικόνα του ιστογράμματος.

Περίληψη Πέντε Αριθμών (Five Number Summary):

- Ελάχιστο (min) = 2844 πόντοι
- Πρώτο τεταρτημόριο (Q1) = 7892 πόντοι
- Διάμεσος (m) = 8398 πόντοι
- Τρίτο τεταρτημόριο (Q3) = 8907 πόντοι
- Μέγιστο (max) = 10371 πόντοι

Η κατανομή των δεδομένων δεν είναι απόλυτα συμμετρική γύρω από τη μέση τιμή, κάτι που αποκαλύπτουν και οι διαφοροποιήσεις μεταξύ του πρώτου και τρίτου τεταρτημορίου. Το γεγονός ότι υπάρχουν ακραίες τιμές (outliers) που αποκλίνουν από τις κεντρικές τιμές, οφείλεται σε μεγάλο βαθμό σε ομάδες που είχαν πολύ διαφορετικές επιδόσεις, όπως για παράδειγμα ομάδες που αποσύρθηκαν ή αγωνίστηκαν σε λιγότερα παιχνίδια μέσα στη σεζόν. Το five-number summary προσφέρει μια πιο λεπτομερή εικόνα για την κατανομή, βοηθώντας στην κατανόηση των διαφοροποιήσεων ανάμεσα στις ομάδες.

Σύγκριση των Μετρικών: Αν και η μέση τιμή και η τυπική απόκλιση προσφέρουν μια αξιόλογη ανασκόπηση της συνολικής τάσης, η σύνοψη των πέντε αριθμών είναι ιδιαίτερα χρήσιμη για τη μελέτη αυτή, ειδικά λόγω της μη συμμετρικής κατανομής των δεδομένων και της ύπαρξης ακραίων τιμών. Η σύνοψη των πέντε αριθμών υποδεικνύει, για παράδειγμα, ότι το 76% των συνολικών πόντων είναι στο εύρος μεταξύ 7892 και 10371 πόντων, ενώ το 24% των παρατηρήσεων είναι κάτω από τους 7892 πόντους.

2. Games Played (G)

Μέση Τιμή (Mean): Η μέση τιμή των αγώνων που έπαιξαν οι ομάδες είναι **79.50** αγώνες. Αυτή η τιμή δίνει μια γενική αίσθηση του μέσου όρου των αγώνων που έπαιξαν οι ομάδες, δηλαδή η πλειοψηφία των ομάδων αγωνίστηκαν περίπου **80** αγώνες σε μια κανονική σεζόν.

Τυπική Απόκλιση (Standard Deviation): Η τυπική απόκλιση είναι **6.5** αγώνες. Αυτό δείχνει ότι οι περισσότερες ομάδες είχαν έναν αριθμό αγώνων που απέκλινε κατά περίπου **6.5** αγώνες από τη μέση τιμή.

Περίληψη Πέντε Αριθμών (Five Number Summary):

- Ελάχιστο (min) = 35 αγώνες
- Πρώτο τεταρτημόριο (Q1) = 82 αγώνες
- Διάμεσος (m) = 82 αγώνες
- Τρίτο τεταρτημόριο (Q3) = 82 αγώνες
- Μέγιστο (max) = 82 αγώνες

Οι σεζόν στο NBA απαρτίζονται από **82 αγώνες**, που αποτελεί τον κανονικό αριθμό αγώνων για κάθε ομάδα στην αρχή της σεζόν. Επομένως, η πλειονότητα των ομάδων θα παίξει **82 αγώνες**, με εξαιρέσεις μόνο για εκείνες που αποσύρθηκαν πρόωρα για διάφορους λόγους.

Αυτή η παρατήρηση μας βοηθά να κατανοήσουμε την κατανομή των δεδομένων. Στο συγκεκριμένο σύνολο δεδομένων, παρατηρούμε ότι η πλειονότητα των ομάδων είτε έπαιξαν το σύνολο των **82 αγώνων** είτε αγωνίστηκαν σε λιγότερους αγώνες λόγω πρόωρης απόσυρσης ή άλλων παραμέτρων. Ωστόσο, η **κατανομή των τιμών** δεν είναι απολύτως συμμετρική γύρω από τη μέση τιμή. Αυτό σημαίνει ότι υπάρχουν **ακραίες τιμές** (outliers) σε περιπτώσεις ομάδων που αποσύρθηκαν νωρίτερα από τον κανονικό αριθμό των αγώνων, κάτι που επηρεάζει την κατανομή των δεδομένων.

Η επιλογή της **σύνοψης των πέντε αριθμών** (Five Number Summary) είναι πιο κατάλληλη για αυτήν την περίπτωση, καθώς παρέχει μια πιο λεπτομερή εικόνα της κατανομής των αγώνων. Με αυτήν τη σύνοψη, μπορούμε να παρατηρήσουμε ότι το **75% των ομάδων** έπαιξαν **82 αγώνες** ή περισσότερους, υποδεικνύοντας ότι η πλειονότητα των ομάδων ολοκλήρωσε τη σεζόν χωρίς πρόωρη απόσυρση. Η **συντομότερη απόσυρση** καταγράφηκε μετά τον **35ο αγώνα**, δηλαδή οι ομάδες που αποσύρθηκαν πρόωρα το έκαναν σχετικά νωρίς στη σεζόν, κάτι που φαίνεται και από την χαμηλή τιμή του **ελάχιστου** στην περίληψη πέντε αριθμών.

Αυτή η ανάλυση μας οδηγεί στο συμπέρασμα ότι, παρόλο που ο κανονικός αριθμός αγώνων είναι **82**, η **προηγούμενη κατανομή** των αγώνων υποδεικνύει την ύπαρξη εξαιρέσεων λόγω των πρόωρων αποσυρμένων ομάδων.

3. Win-Loss Ration (W.L.)

Μέση Τιμή (Mean): Η **μέση τιμή** των δεδομένων είναι **0.499**. Αυτό δείχνει ότι η μέση τιμή των παρατηρήσεων βρίσκεται περίπου στο **0.50**, κάτι που υποδεικνύει ότι οι παρατηρήσεις είναι συγκεντρωμένες γύρω από αυτή την τιμή.

Τυπική Απόκλιση (Standard Deviation): Η **τυπική απόκλιση** είναι **0.145**. Αυτό σημαίνει ότι οι παρατηρήσεις παρουσιάζουν μέτρια διασπορά γύρω από τη μέση τιμή, με τις περισσότερες παρατηρήσεις να απέχουν κατά περίπου **0.145** από τον μέσο όρο.

Περίληψη Πέντε Αριθμών (Five Number Summary):

- Ελάχιστο (min) = 0,106
- Πρώτο τεταρτημόριο (Q1) = 0,390
- Διάμεσος (m) = 0,512
- Τρίτο τεταρτημόριο (Q3) = 0,610
- Μέγιστο (max) = 0,890

Όταν η κατανομή μιας μεταβλητής ακολουθεί **κανονική κατανομή**, όπως φαίνεται από το ιστόγραμμα και το QQ-plot, σημαίνει ότι οι τιμές είναι **συμμετρικά κατανεμημένες γύρω από τη μέση τιμή**. Στην περίπτωση αυτή, η **μέση τιμή (*mean*)** και η **τυπική απόκλιση (*standard deviation*)** είναι τα πιο κατάλληλα στατιστικά μέτρα για να περιγράψουν την κατανομή και τη διακύμανση των δεδομένων.

Επειδή η κατανομή των **W to L ratio** είναι συμμετρική και ακολουθεί κανονική κατανομή, τα στατιστικά μέτρα όπως η μέση τιμή και η τυπική απόκλιση παρέχουν αξιόπιστες και χρήσιμες πληροφορίες για την περιγραφή των δεδομένων και την κατανόηση της κατανομής τους.

e) **Διερεύνηση σχέσης Games Played-Points Scored**

Η διερεύνηση της σχέσης μεταξύ των μεταβλητών **Games Played (*G*)** και **Points Scored (*PTS*)** είναι σημαντική για να κατανοήσουμε κατά πόσο η ποσότητα των αγώνων που παίζει μια ομάδα σε μια σεζόν επηρεάζει τους πόντους που πετυχαίνει.

1. **Σκοπός και Επιλογή Μεταβλητών:**

Η μεταβλητή **G (*Games Played*)** θα θεωρηθεί η **επεξηγηματική μεταβλητή** (*independent variable*), ενώ η μεταβλητή **PTS (*Points Scored*)** θα είναι η **μεταβλητή απόκρισης** (*dependent variable*). Δηλαδή, προσπαθούμε να καταλάβουμε αν και πώς οι αγώνες που παίζει μια ομάδα (*G*) επηρεάζουν τους πόντους που πετυχαίνει (*PTS*).

2. Ανάλυση μέσω Scatter Plot:

Από το **scatter plot** μπορούμε να διακρίνουμε αν υπάρχει σχέση μεταξύ των δύο μεταβλητών και να καταλάβουμε αν αυτή η σχέση είναι γραμμική ή μη γραμμική.

- **Γραμμική σχέση** σημαίνει ότι η αύξηση της τιμής της **G** οδηγεί σε μια αντίστοιχη αύξηση ή μείωση της τιμής της **PTS** με σταθερό ρυθμό.
- **Μη γραμμική σχέση** υποδηλώνει ότι η σχέση είναι πιο περίπλοκη, δηλαδή, η αύξηση των αγώνων δεν οδηγεί σε μια συνεχή ή σταθερή αλλαγή στους πόντους.

Όταν λέμε ότι το scatter plot δείχνει μια **αύξουσα σχέση**, αυτό σημαίνει ότι καθώς οι αγώνες αυξάνονται, οι πόντοι τείνουν να αυξάνονται, αλλά η σχέση δεν είναι αυστηρά γραμμική (*δηλαδή μπορεί να υπάρχουν κάποιες παραλλαγές στη συμπεριφορά των δεδομένων*).

3. Συντελεστής Συσχέτισης (Correlation Coefficient):

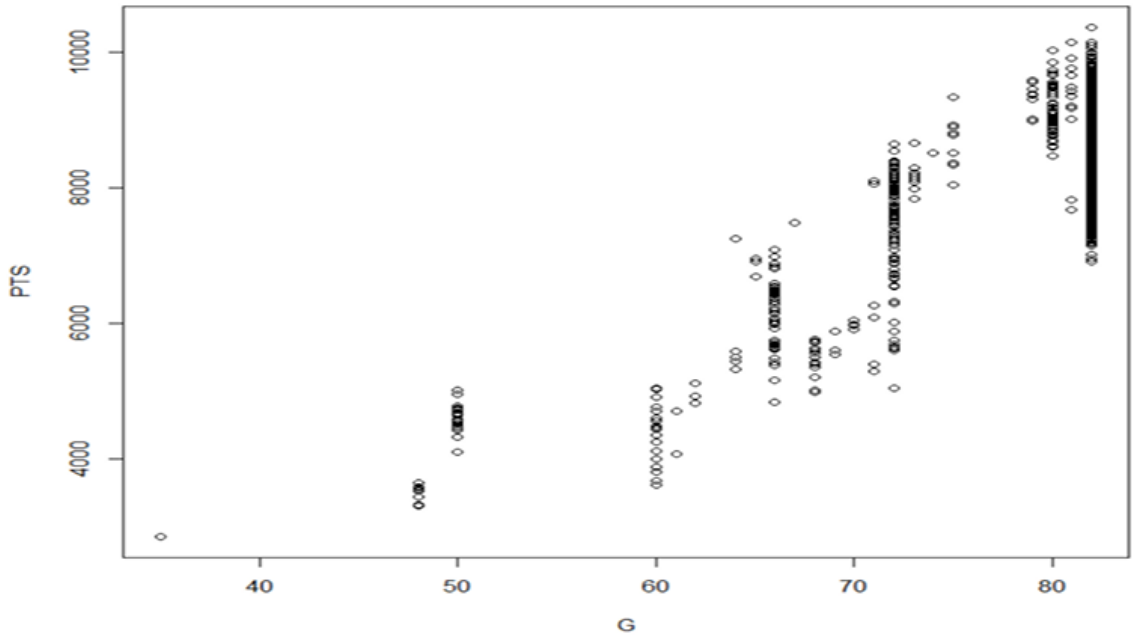
Για να ποσοτικοποιήσουμε τη σχέση μεταξύ των δύο μεταβλητών, μπορούμε να υπολογίσουμε τον **συντελεστή συσχέτισης**, ο οποίος μας δείχνει πόσο ισχυρή και σε ποια κατεύθυνση είναι η σχέση μεταξύ τους.

Ο συντελεστής συσχέτισης **r** κυμαίνεται μεταξύ **-1 και 1**:

- **r = 1**: Απόλυτη θετική γραμμική συσχέτιση (*δηλαδή, με την αύξηση της μίας μεταβλητής αυξάνεται η άλλη με σταθερό ρυθμό*).
- **r = -1**: Απόλυτη αρνητική γραμμική συσχέτιση (*όταν η μία μεταβλητή αυξάνεται, η άλλη μειώνεται με σταθερό ρυθμό*).
- **r = 0**: Καμία γραμμική συσχέτιση.
- **0 < r < 1**: Υπάρχει θετική συσχέτιση, αλλά δεν είναι απόλυτη (*όσο πιο κοντά στο 1, τόσο πιο ισχυρή είναι η θετική γραμμική συσχέτιση*).
- **-1 < r < 0**: Υπάρχει αρνητική συσχέτιση, αλλά δεν είναι απόλυτη.

Με αυτόν τον τρόπο, υπολογίζοντας τον συντελεστή συσχέτισης, μπορούμε να καταλάβουμε πόσο ισχυρή είναι η σχέση μεταξύ των **Games Played (*G*)** και **Points Scored (*PTS*)**.

Από το **scatterplot** μπορούμε να παρατηρήσουμε μια **αύξουσα** αλλά **όχι απόλυτα γραμμική σχέση** μεταξύ των δύο μεταβλητών: **Games Played (*G*)** και **Points Scored (*PTS*)**. Η σχέση αυτή, αν και είναι γενικά θετική (*δηλαδή, όσο περισσότερους αγώνες παίζει μια ομάδα, τόσο περισσότερους πόντους πετυχαίνει*), δεν ακολουθεί μια τέλεια ευθεία γραμμή. Αυτό υποδεικνύει ότι η σχέση μπορεί να είναι **μη γραμμική** ή απλά ότι υπάρχουν άλλοι παράγοντες που επηρεάζουν την αύξηση των πόντων πέρα από τον αριθμό των αγώνων. Ωστόσο, για να κατανοήσουμε καλύτερα τη δύναμη και τη φύση αυτής της σχέσης, το επόμενο λογικό βήμα είναι να υπολογίσουμε τον **συντελεστή συσχέτισης** (*r*). Ο συντελεστής συσχέτισης μπορεί να μας δώσει μια ποσοτική εκτίμηση για το πόσο ισχυρή είναι η σχέση και αν αυτή είναι θετική ή αρνητική.



Εντολές σε R

```
> cor(G, PTS)
```

Ο συντελεστής **r = 0.8179445** υποδηλώνει μια **ισχυρή θετική γραμμική συσχέτιση**, πράγμα που σημαίνει ότι καθώς αυξάνονται οι αγώνες (*Games Played*), αυξάνονται επίσης και οι συνολικοί πόντοι (*Points Scored*) με σχετικά υψηλή αξιοπιστία. Με άλλα λόγια, όσο περισσότεροι αγώνες παίζει μια ομάδα, τόσο περισσότερους πόντους πετυχαίνει στην διάρκεια της σεζόν, και αυτή η σχέση είναι αρκετά ισχυρή.

Η σχέση αυτή δεν είναι απλώς συσχέτιση, αλλά **αιτιατή**. Δηλαδή, η μεταβολή στον αριθμό των αγώνων (*G*) προκαλεί σίγουρη μεταβολή στους συνολικούς πόντους (*PTS*). Επειδή οι ομάδες του NBA σκοράρουν κατά μέσο όρο από **90 μέχρι 110 πόντους ανά παιχνίδι**, το γεγονός ότι μια ομάδα δεν αγωνίζεται σε ένα ή περισσότερους αγώνες έχει σημαντική επίπτωση στους συνολικούς πόντους της σεζόν. Για παράδειγμα:

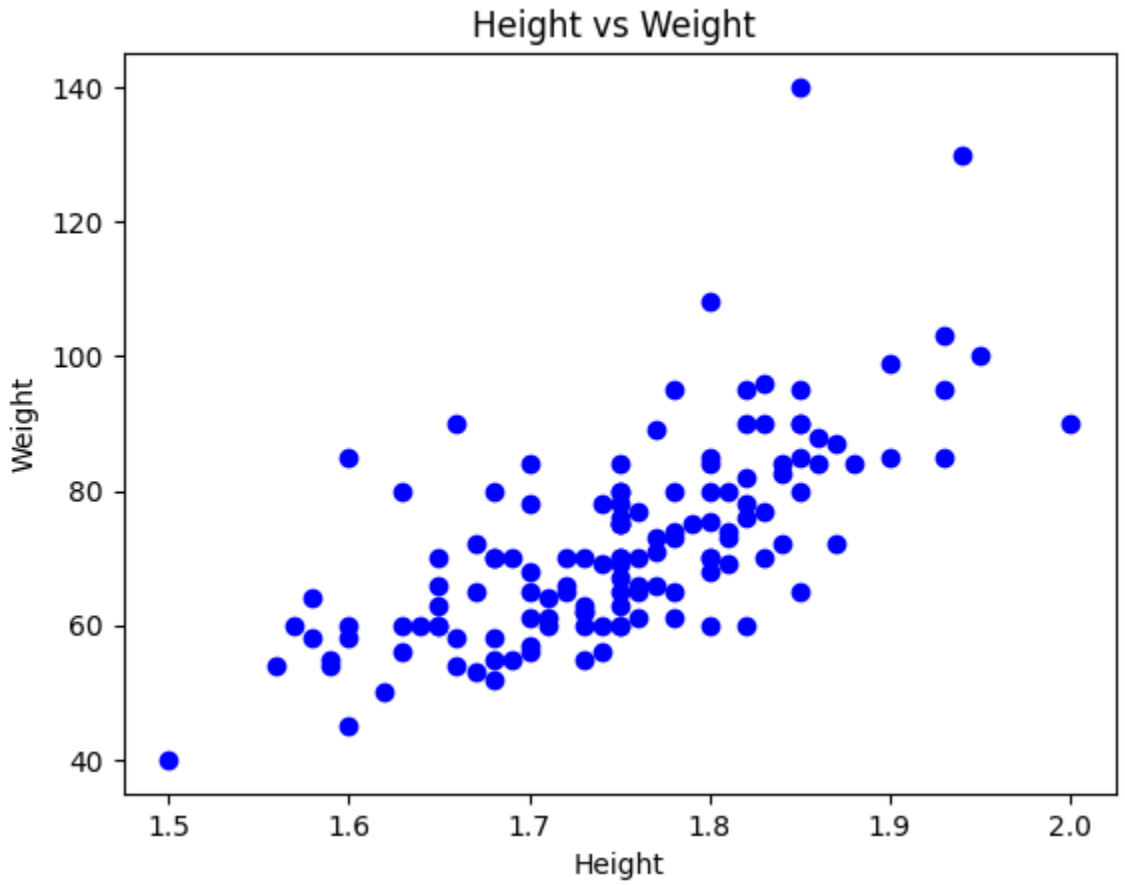
- Αν μια ομάδα δεν αγωνιστεί σε έναν ή δύο αγώνες, αυτό μπορεί να μειώσει κατά πολύ τον συνολικό αριθμό πόντων που θα πετύχει στο τέλος της σεζόν.
- Αντίθετα, όσο περισσότεροι αγώνες παίζει μια ομάδα, τόσο πιο πολλούς πόντους θα πετύχει λόγω της αυξημένης ευκαιρίας για σκοράρισμα.

Η σχέση αυτή δείχνει πόσο σημαντικός είναι ο αριθμός των αγώνων για τον συνολικό αριθμό πόντων της ομάδας. Επομένως, η συνέχιση των αγώνων έχει σημαντική αιτιώδη σχέση με την απόδοση της ομάδας, κάτι που επιβεβαιώνεται από την ισχυρή συσχέτιση. Η αλλαγή στο G (*π.χ. αν μια ομάδα αποσύρεται ή δεν συμμετέχει σε αγώνες*) μπορεί να έχει πολύ σοβαρές συνέπειες στο τελικό σύνολο των πόντων (*PTS*), κάτι που καθιστά τον αριθμό των αγώνων κρίσιμο για την επιτυχία της ομάδας.

Άσκηση 3

a) Για να διερευνήσουμε τη σχέση μεταξύ του **ύψους (*height*)** και του **βάρους (*weight*)** των ατόμων για το 2024, επιλέξαμε αυτές τις δύο ποσοτικές μεταβλητές και κατασκευάσαμε ένα **scatterplot**. Στο scatterplot, το **ύψος** τοποθετείται στον **άξονα x** και το **βάρος** στον **άξονα y**, καθώς θεωρούμε το ύψος ως επεξηγηματική μεταβλητή (*ή ανεξάρτητη μεταβλητή*) και το βάρος ως μεταβλητή απόκρισης (*ή εξαρτημένη μεταβλητή*).

Αυτή η επιλογή γίνεται με βάση την παραδοχή ότι το ύψος ενός ατόμου μπορεί να επηρεάσει το βάρος του, δηλαδή η αύξηση στο ύψος ενδέχεται να σχετίζεται με την αύξηση στο βάρος.



Μορφή της σχέσης: Η σχέση φαίνεται να είναι **γραμμική**, καθώς τα σημεία σχηματίζουν μια διαγώνια, περίπου ευθύγραμμη κατανομή. Αυτό υποδεικνύει ότι το βάρος (*weight*) αυξάνεται αναλογικά με το ύψος (*height*).

Κατεύθυνση της σχέσης: Η κατεύθυνση της σχέσης είναι **θετική(αύξουσα)**, δηλαδή όσο αυξάνεται το ύψος, τείνει να αυξάνεται και το βάρος. Αυτό είναι εμφανές από την ανοδική τάση των σημείων.

Δύναμη της σχέσης: Η δύναμη της σχέσης είναι **θετική και ισχυρή**. Η συσχέτιση είναι ισχυρή, καθώς τα περισσότερα σημεία ακολουθούν έναν σαφή γραμμικό κανόνα. Η θετική αυτή συσχέτιση φαίνεται καθαρά από το συγκεντρωμένο μοτίβο των σημείων.

b) Ο συντελεστής συσχέτισης μεταξύ ύψους (*height*) και βάρους (*weight*) είναι **0.6811889**, κάτι που υποδηλώνει μια **ισχυρή θετική συσχέτιση**. Αυτό σημαίνει ότι, καθώς αυξάνεται το ύψος, υπάρχει γενικά μια τάση για αύξηση του βάρους, κάτι που συνάδει με την αρχική μας παρατήρηση από το διάγραμμα.