Στατιστική Στην Πληροφορική - 1η Άσκηση
ΟΠΑ , Ακαδημαϊκό Έτος: 2020-2021

Ομάδα #0

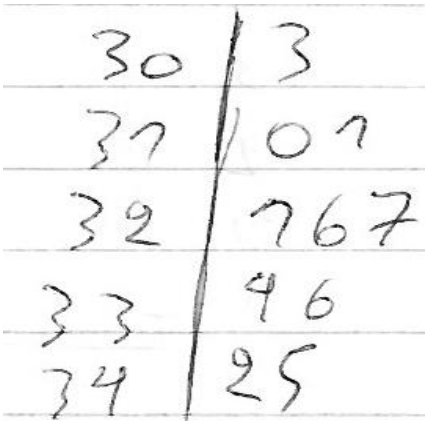
- ΤΣΙΟΜΠΙΚΑΣ ΔΗΜΗΤΡΙΟΣ , 3180223
- ΠΑΝΑΓΙΩΤΟΥ ΠΑΝΑΓΙΩΤΗΣ , 3180139

1ο Ερώτημα

a. Ακολουθούν τα **stemplots** και τα **boxplots** των δεδομένων που δόθηκαν στην εκφώνηση. (Οι ασκήσεις λύθηκαν πρώτα σε χαρτί. Δε χρησιμοποιήθηκε software για τη επίλυση τους)

- Για τα Δεδομένα Ι

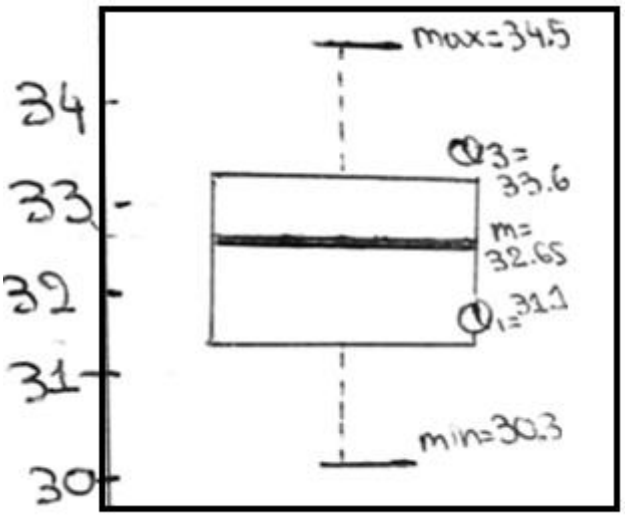
Stemplot



BoxPlot

Η *σύνοψη των 5 αριθμών* είναι:

- **Min** = 30.3
- **Q1** = 31.1
- **M** = (32.6+32.7)/2=32.65
- **Q3** = 33.6
- **Max** = 34.5
- **[Κάτω Φράγμα, Άνω Φράγμα]**= [Q1-1.5IQR, Q3+1.5IQR]= [27.35 , 37.35]
  Άρα δεν υπάρχουν ατυπικές τιμές.



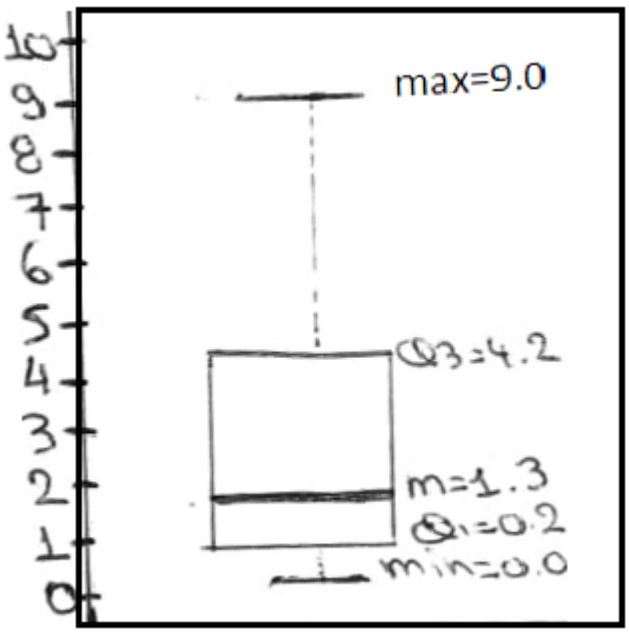
- <u>Για τα Δεδομένα ΙΙ</u>

Stemplot

BoxPlot

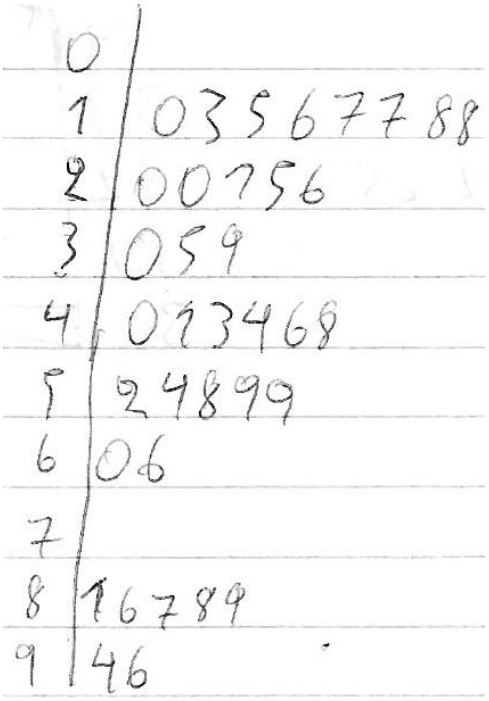
Η *σύνοψη των 5 αριθμών* είναι:

- **Min** = 0.0
- **Q1** = 0.2
- **M** = (1.2+1.4)/2=1.3
- **Q3** = 4.2
- **Max** = 9.0
- **[Κάτω Φράγμα, Άνω Φράγμα]**= [Q1-1.5IQR, Q3+1.5IQR]= [-5.8 , 10.2]
  Άρα <u>δεν υπάρχουν ατυπικές τιμές.</u>



- <u>Για τα Δεδομένα ΙΙΙ</u>

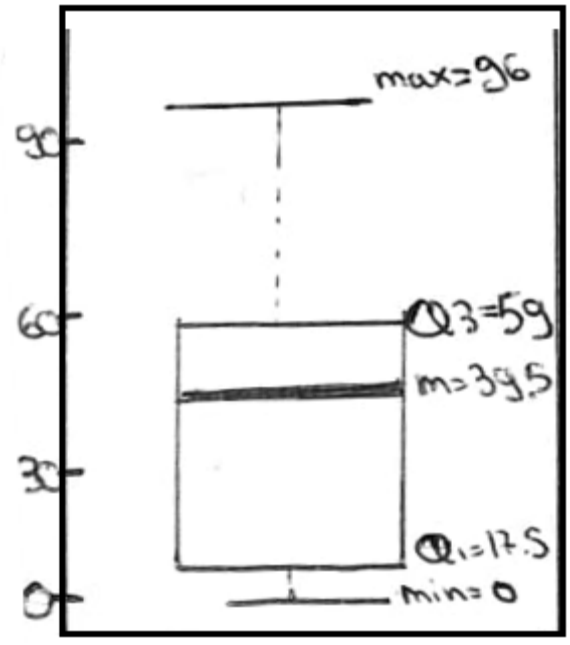
Stemplot



BoxPlot

Η *σύνοψη των 5 αριθμών* είναι:

- **Min** = 0
- **Q1** = (17+18)/2= 17,5
- **M** = (39+40)/2= 39.5
- **Q3** = (59+59)/2 = 59
- **Max** = 96
- **[Κάτω Φράγμα, Άνω Φράγμα]**= [Q1-1.5IQR, Q3+1.5IQR]= [-44.75 , 121.25]
  Άρα <u>δεν υπάρχουν ατυπικές τιμές.</u>



b.

- Για τα Δεδομένα I

Τυπική απόκλιση = 1.419898

Μέση τιμή = 32.55

Σε αυτά τα δεδομένα και οι 2 τρόποι είναι επαρκείς για να συνοψίσουν την κατανομή. Εδώ θα μπορούσαμε να θεωρήσουμε λίγο καλύτερο τον συνδυασμό **μέσης τιμής – τυπικής απόκλισης** καθώς η μέση τιμή είναι πολύ κοντά στη διάμεση τιμή, η τυπική απόκλιση είναι μικρή και η κατανομή είναι ομοιόμορφη.

- Για τα Δεδομένα II

Τυπική απόκλιση = 3.059121

Μέση τιμή = 2.64

Σε αυτά τα δεδομένα θα ήταν καλύτερο να χρησιμοποιήσουμε το **five numbers summary** (*boxplot*) για την σύνοψη της κατανομής καθώς είναι προφανές ότι υπάρχει μία συγκέντρωση των δεδομένων προς τα κάτω και μία αραίωση προς τα πάνω (μη συμμετρική κατανομή) μέσω του boxplot , η μέση τιμή είναι μικρότερη της τυπικής απόκλισης (2.64 < 3.059121) και απέχει λίγο από τη διάμεση τιμή.

- Για τα Δεδομένα III

Τυπική απόκλιση = 28.29804

Μέση τιμή = 41.125

Σε αυτά τα δεδομένα είναι καλύτερη η χρήση για άλλη μια φορά του **five number summary** (*boxplot*) διότι έχουμε πάλι συγκέντρωση των δεδομένων προς τα κάτω (μη συμμετρική κατανομή) και μια αραίωση προς τα πάνω το οποίο το διαπιστώνουμε μέσω του boxplot.

c. Ξεκινάμε με βάση τον ***αλγόριθμο περιγραφής ποσοτικής*** **μεταβλητής** με καμπύλη πυκνότητας.

- Για τα Δεδομένα Ι

-Διάταξη σε αύξουσα σειρά :

Έχουμε πλήθος δεδομένων N = 10, άρα $X_1 \leq X_2 \leq X_3 \ldots \leq X_{10}$

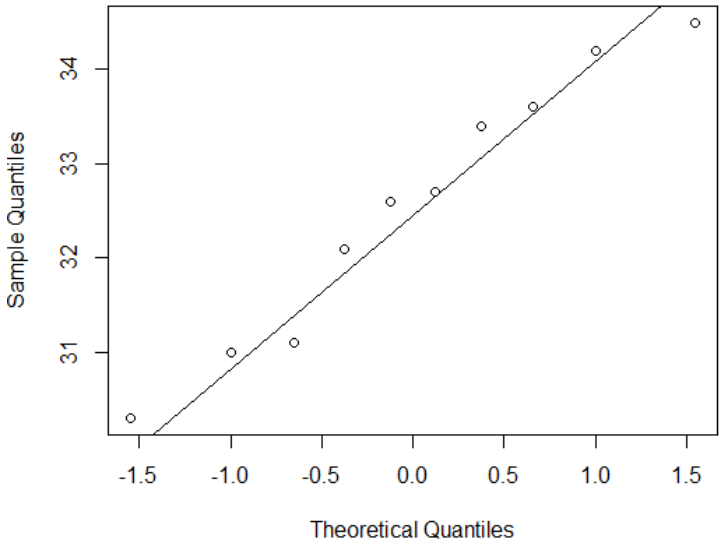
$X_1 = 30$
$X_2 = 31$
$X_3 = 31$
$X_4 = 32$
$X_5 = 32$
$X_6 = 32$
$X_7 = 33$
$X_8 = 33$
$X_9 = 34$
$X_{10} = 34$

-Εύρεση % τιμών αριστερά κάθε τιμής :

$P_1 = 0\%$
$P_2 = 10\%$
$P_3 = 20\%$
$P_4 = 30\%$
$P_5 = 40\%$
$P_6 = 50\%$
$P_7 = 60\%$
$P_8 = 70\%$
$P_9 = 80\%$
$P_{10} = 90\%$

-Υπολογισμός των $X_1 <= X_2 <= X_3 \ldots <= X_{10}$ όπου $X_1 = P_1$, $X_2 = P_2$,.. $X_{10} = P_{10}$

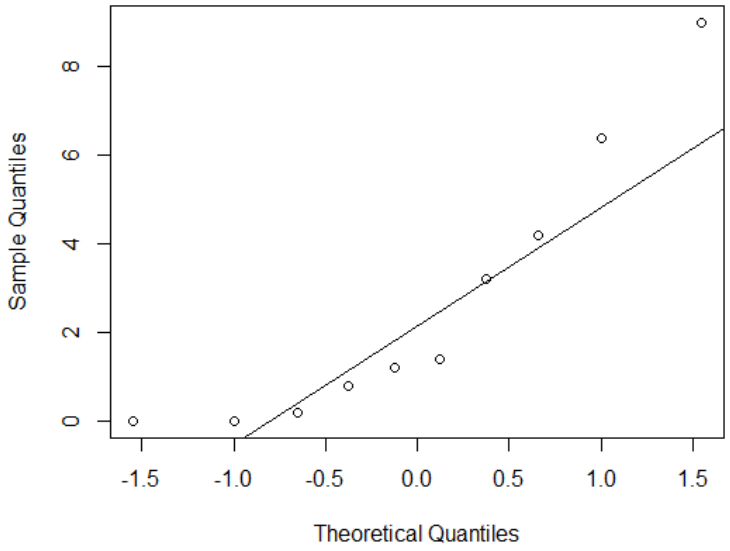
-Τέλος, το Normal-Quantile plot:



Εδώ παρατηρούμε ότι όλα τα στοιχεία είναι πολύ κοντά στη γραμμή χωρίς τρομακτικές αποκλίσεις. Άρα **τα Δεδομένα Ι προσεγγίζουν επαρκώς την κανονική κατανομή**.

- Για τα Δεδομένα II

Κάνουμε την ίδια διαδικασία και εδώ και το Normal Quantile plot είναι το εξής :



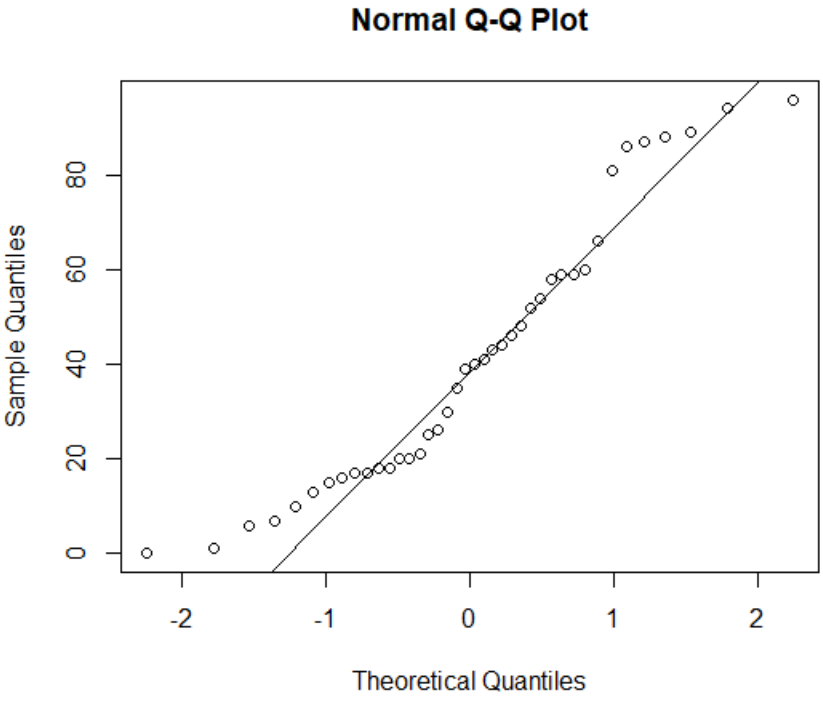
Στα δεδομένα αυτά τα περισσότερα στοιχεία είναι κοντά στη γραμμή. Όμως δημιουργούνται μερικές μικρές καμπύλες λόγω ορισμένων στοιχείων που είναι αρκετά απομακρυσμένα από τη γραμμή, συγκεκριμένα στο χαμηλότερο τεταρτημόριο και στο υψηλότερο τεταρτημόριο. Άρα **τα Δεδομένα II δεν προσεγγίζουν επαρκώς την κανονική κατανομή**.

Ο κώδικας που γράφτηκε (και με παρόμοιο τρόπο για τα υπόλοιπα δεδομένα) ήταν:

```
> y<-c(0.0 , 0.0 , 0.2 , 0.8 , 1.2 , 1.4 , 3.2 , 4.2 , 6.4 , 9.0)
> qqnorm(y)
> qqline(y)
```

- Για τα Δεδομένα III

Για άλλη μία φορά εκτελούμε την ίδια διαδικασία. Το Normal Quantile plot είναι το εξής :



Εδώ παρατηρούμε ότι τα Δεδομένα III είναι μεγάλα σε πλήθος και έχουμε μια καλύτερη εικόνα γενικά για την προσέγγιση. Αρχικά, τα περισσότερα δεδομένα προσεγγίζουν τη γραμμή αλλά πάλι βλέπουμε μια ανωμαλία στα χαμηλά ποσοστημόρια (0%-25%) και στα ψηλά ποσοστημόρια (60%-80%). Εκεί έχουμε απόκλιση των στοιχείων από τη γραμμή και εμφάνιση καμπυλών που καθιστούν το σχήμα ελικοειδές.

Βάσει όλων αυτών των παρατηρήσεων, συμπεραίνουμε ότι **τα Δεδομένα III δεν προσεγγίζουν κατάλληλα την κανονική κατανομή**.

2ο Ερώτημα

a. Τα δεδομένα προέρχονται από την ιστοσελίδα Serebii (https://serebii.net/), μια από τις πιο αξιόπιστες databases για ότι αφορά τα δεδομένα των Pokémon βιντεοπαιχνιδιών και του franchise γενικότερα.

Επιλέξαμε ως δεδομένα τα πρώτα 151 Pokémon (1st Generation), όπου το καθένα έχει τα δικά του μοναδικά χαρακτηριστικά στοιχεία (όνομα,είδος,αριθμός,ταχύτητα κλπ). Τα δεδομένα αφορούν μόνο τα πρώτα Pokémon παιχνίδια (Red ,Green,Blue,Yellow) των 90s.

b. Διάκριση των μεταβλητών σε κατηγορικές και ποσοτικές:

Κατηγορικές:

- **Number, Name**: Κάθε Pokémon έχει το δικό του μοναδικό αριθμό (id) και όνομα.
- **Type1, Type2**: Κάθε Pokémon μπορεί να ανήκει μέχρι και σε δύο τύπους. Πχ: Fire και Flying. Οι τύποι σχετίζονται με τη φύση που ζουν και τη φυσική τους κατάσταση. (πχ. Ένα Pokémon που ζει στο νερό θα είναι σίγουρα Water type).
- **Legendary**: Ένα Pokémon θεωρείται θρυλικό μόνο αν είναι σπάνιο ή πολύ δυνατό.

Ποσοτικές:

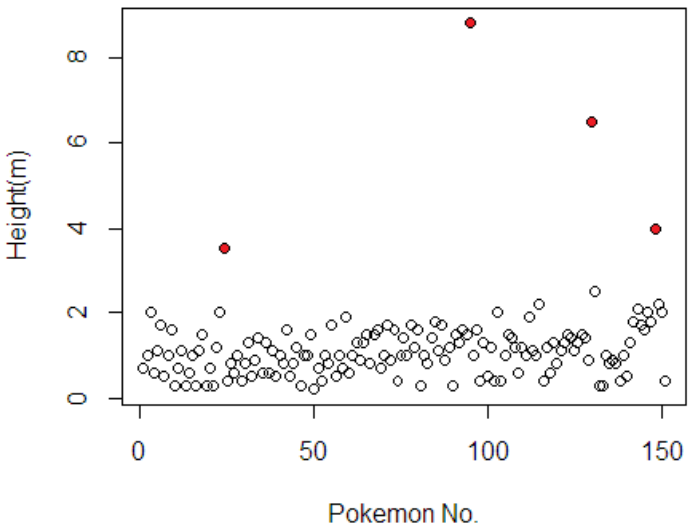
- **Height, Weight**: Περιγράφουν το ύψος και το βάρος τους.
- **HP, Attack, Defense, Speed:** Οι μεταβλητές αυτές αφορούν τις βασική «δύναμη» των Pokémon. Γνωστά και ως “Base Stats”. Όσο μεγαλύτερη είναι η κάθε μία από αυτές μεταβλητή, τόσο πιο πολύ θα δυναμώνουν τα Pokémon σε αυτό το “Stat” καθώς ανεβαίνουν επίπεδο. Το HP (Hit Points) αφορά τη «ζωή» τους στη μάχη. Το Attack αφορά τη δύναμη στις επιθέσεις που κάνει το Pokémon προς τον αντίπαλο , το Defense την άμυνα κατά των επιθέσεων του αντιπάλου και το Speed το πόσο γρήγορο είναι (δηλαδή αν αυτό θα καταφέρει να επιτεθεί πρώτο σε ένα γύρο).
- **Evolutions:** Η μεταβλητή αυτή μετράει πόσες εξελίξεις έχει ένα Pokémon μέχρι να φτάσει στην τελική του μορφή.

c. Γράψαμε τον παρακάτω κώδικα στην R:

```
# Κάνουμε import το .csv αρχείο στην R
> read.csv("C:\\Users\\peter\\Desktop\\FirstGenPokemon.csv", header = TRUE) -> TempTable
# Αφαιρούμε τις στήλες με τις οποίες δε θα ασχοληθούμε στην έρευνα αυτή
> MyTable = subset(TempTable, select = -c(Types,Capt_Rate, Male_Pct, Female_Pct,Exp_Points, Exp_Speed,
Base_Total,Special,Normal_Dmg, Fire_Dmg, Water_Dmg, Eletric_Dmg, Grass_Dmg, Ice_Dmg, Fight_Dmg,
Poison_Dmg, Ground_Dmg, Flying_Dmg, Psychic_Dmg, Bug_Dmg, Rock_Dmg, Ghost_Dmg, Dragon_Dmg))
# Ότι χρειαζόμαστε αυτή τη στιγμή βρίσκεται στον πίνακα MyTable
```

Στον οριζόντια άξονα θα έχουμε σταθερά τον αριθμό των Pokémon (δηλαδή 1 – 151).

Στον κατακόρυφο άξονα θα έχουμε κάθε φορά τη μεταβλητή της οποίας εξετάζουμε την κατανομή.



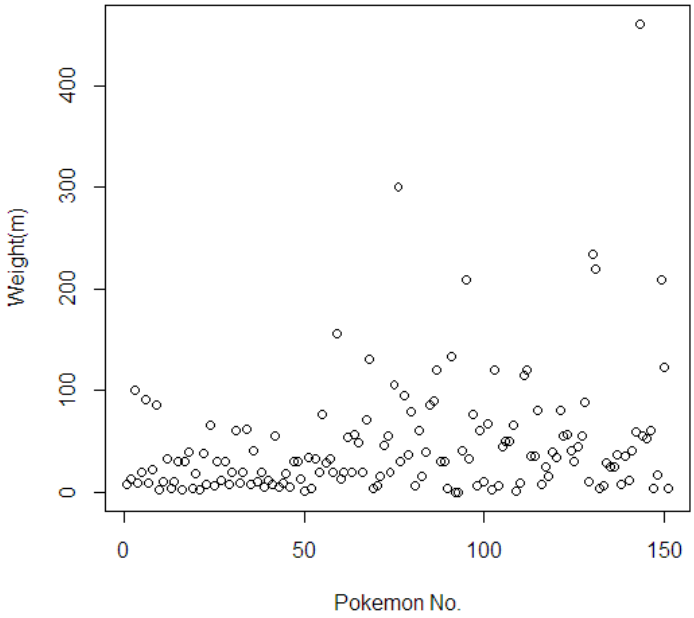
Height.m. : Παρατηρούμε πως υπάρχουν 4 ατυπικές τιμές (σημειώσαμε εμείς με κόκκινο). Δηλαδή υπάρχουν 4 Pokémon με μεγάλη διαφορά ύψους/μήκους από τα υπόλοιπα. Πιο συγκεκριμένα:

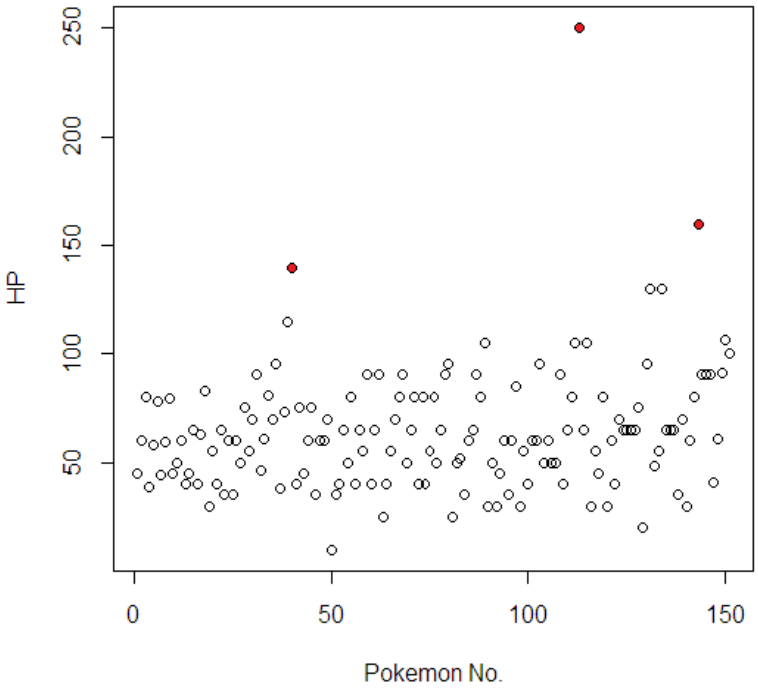
1. **No.95** , **Onix** με 8.8m

(Ακόμα και σήμερα, μετά από 8 γενιές, βρίσκεται στα Top 10 πιο ψηλά Pokémon)

2. **No.130** , **Gyarados** με 6.5m
3. **No.148** , *Dragonair* με 4.0m
4. **No.24**, **Arbok** με 3.5m

Weight.kg. : Με μία πρώτη ματιά παρατηρούμε πως υπάρχουν αρκετές ατυπικές τιμές (τουλάχιστον πάνω από 5). Το μεγαλύτερο βάρος με διαφορά όμως το έχει το **No.143, Snorlax** με βάρος 460kg. Το Snorlax είναι ένα από τα πιο χαρακτηριστικά και διάσημα Pokémon για αυτό το λόγο. Πολύς κόσμος χωρίς να είναι θαυμαστής του franchise γνωρίζει πως μοιάζει εμφανισιακά.





HP. : Παρατηρούμε πως υπάρχουν 3 ατυπικές τιμές (σημειώσαμε εμείς με κόκκινο). Δηλαδή υπάρχουν 3 Pokémon με περισσότερη ικανότητα στο HP τους καθώς κάνουν level-up. Το HP (όπως και το Attack,Defense και Speed) αυξάνεται καθώς το Pokémon ανεβαίνει επίπεδα. Ένα Pokémon με υψηλό base HP stat θα αυξάνει το HP του με μεγάλους ρυθμούς (π.χ +8 ανά επίπεδο, ενώ άλλα με χαμηλό μόνο +2).

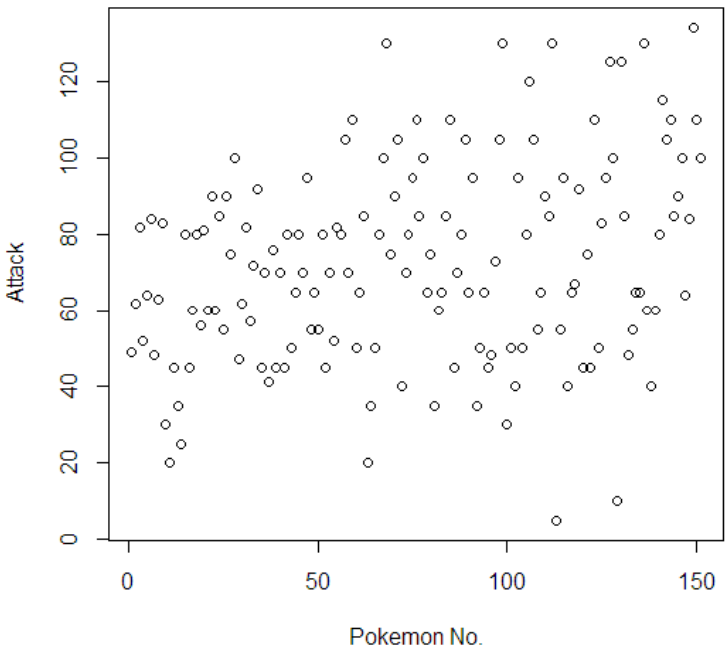
Τα 3 Pokémon με το υψηλότερο HP stat είναι:

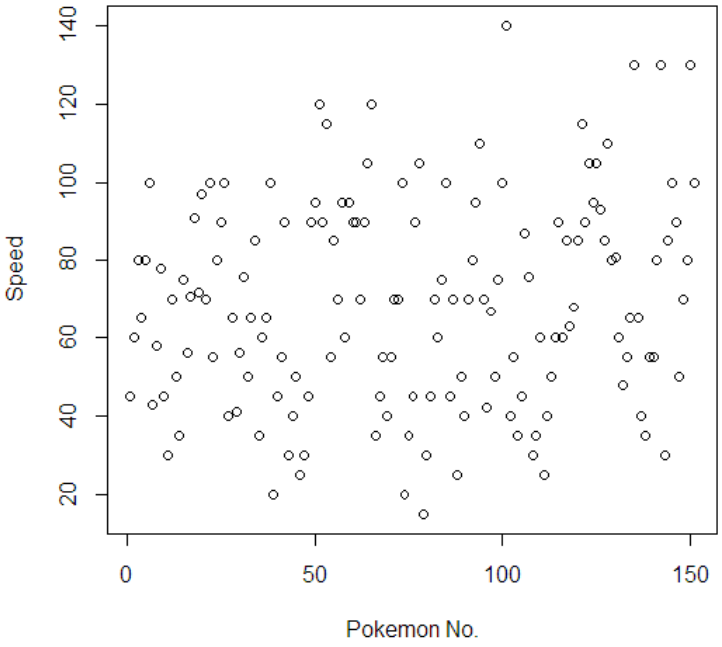
1. **No. 113 , Chansey** με 250 (Ξεπερνάει κατά πολύ όλα τα υπόλοιπα καθώς έχει καθιερωθεί ως το το Pokémon της «υγείας»)

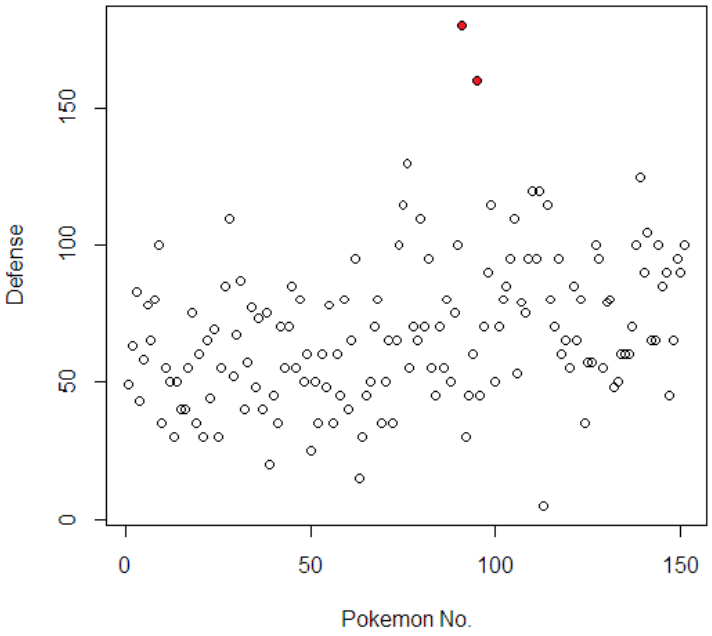
2. **No. 143 , Snorlax** με 160

3. **No. 40 , Wigglytuff** με 140

Παρακάτω βλέπουμε τις κατανομές των μεταβλητών **Attack** και **Speed** που είναι αρκετά παρόμοιες στην όψη. Δεν υπάρχουν ατυπικές τιμές και η κατανομή φαίνεται αρκετά ομοιόμορφη και στις δύο.







Defense: Παρατηρούμε πως υπάρχουν 2 ατυπικές τιμές (σημειώσαμε εμείς με κόκκινο). Πιο συγκεκριμένα αφορούν τα εξής Pokémon :

1. **No 91. Cloyster** με 180

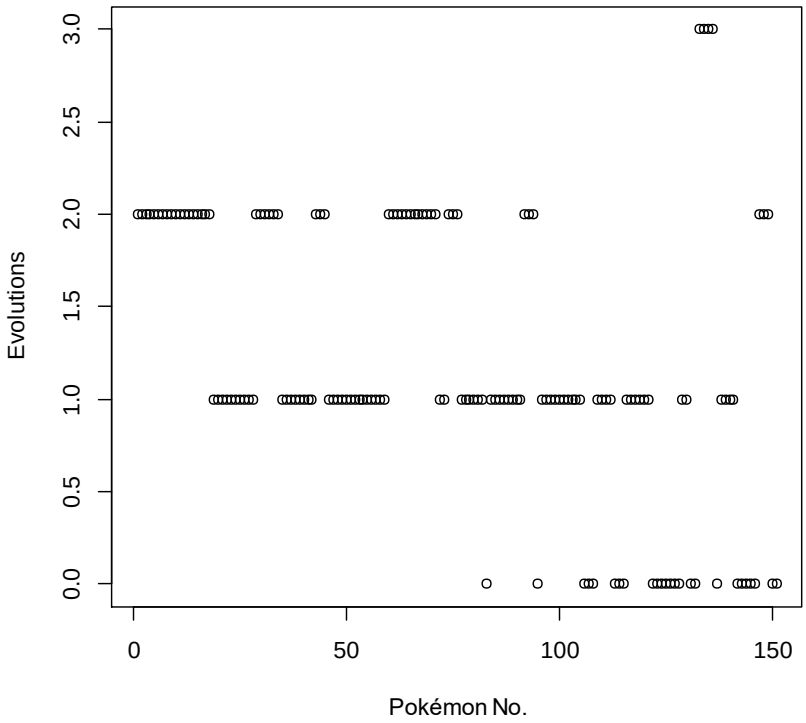
(Το σώμα του αποτελείται από ένα σκληρό καβούκι με καρφιά για αυτό έχει τόσο προστασία)

2. **No 95. Onix** με 160

(Το σώμα του αποτελείται από πέτρα)

Evolutions: Με μια ματιά βλέπουμε πως τα περισσότερα Pokémon έχουν 1 μόνο εξέλιξη. Πολλά από τα υπόλοιπα έχουν 2 εξελίξεις και αρκετά είναι αυτά που δεν έχουν καθόλου εξελίξεις.

Πολύ λίγα, συγκεκριμένα 4, είναι αυτά που έχουν 3 εξελίξεις. Συγκεκρίμενα τα **Eevee,Vaporeon, Jolteon, Flareon** . Ανήκουν όλα στο ίδιο “Evolution Family”, καθώς το Eevee μπορεί να εξελιχθεί και στα 3 από αυτά. Το Eevee είναι και γνωστό ως το “Evolution Pokémon” (στις επόμενες γενιές αποκτά άλλες 5 εξελίξεις).



d.

α. Χρησιμοποιούμε την παρακάτω συνάρτηση για να πάρουμε την **τυπική απόκλιση:**

```
> sd(MyTable$ΌνομαΜεταβλητής)
```

Χρησιμοποιούμε την παρακάτω συνάρτηση για να πάρουμε τη **μέση τιμή**:

```
> mean(MyTable$ΌνομαΜεταβλητής)
```

Παίρνουμε τα παρακάτω αποτέλεσματα.

```
> sd(MyTable$Height.m.)
[1] 0.9626206
> sd(MyTable$Weight.kg.)
[1] 59.44799
> sd(MyTable$HP)
[1] 28.59012
> sd(MyTable$Attack)
[1] 26.31372
> sd(MyTable$Defense)
[1] 26.99879
> sd(MyTable$Speed)
[1] 26.58872
> sd(MyTable$Evolutions)
[1] 0.7422298
```

```
> mean(MyTable$Height.m.)
[1] 1.194702
> mean(MyTable$Weight.kg.)
[1] 45.95166
> mean(MyTable$HP)
[1] 64.21192
> mean(MyTable$Attack)
[1] 72.21854
> mean(MyTable$Defense)
[1] 68.15894
> mean(MyTable$Speed)
[1] 68.80132
> mean(MyTable$Evolutions)
[1] 1.205298
```

β. Χρησιμοποιούμε την παρακάτω συνάρτηση για να πάρουμε την **σύνοψη των 5 αριθών**:

```
> summary(MyTable)
```

Παίρνουμε το αποτέλεσμα της συνάρτησης (κρύψαμε τις κατηγορικές μεταβλητές)

```
   Height.m.       Weight.kg.          HP             Attack          Defense          Speed         Evolutions   
Min.   :0.200   Min.   :  0.10   Min.   : 10.00   Min.   :  5.00   Min.   :  5.00   Min.   : 15.0   Min.   :0.000  
1st Qu.:0.700   1st Qu.:  9.90   1st Qu.: 45.00   1st Qu.: 51.00   1st Qu.: 50.00   1st Qu.: 46.5   1st Qu.:1.000  
Median :1.000   Median : 30.00   Median : 60.00   Median : 70.00   Median : 65.00   Median : 70.0   Median :1.000  
Mean   :1.195   Mean   : 45.95   Mean   : 64.21   Mean   : 72.22   Mean   : 68.16   Mean   : 68.8   Mean   :1.205  
3rd Qu.:1.500   3rd Qu.: 56.25   3rd Qu.: 80.00   3rd Qu.: 90.00   3rd Qu.: 84.00   3rd Qu.: 90.0   3rd Qu.:2.000  
Max.   :8.800   Max.   :460.00   Max.   :250.00   Max.   :134.00   Max.   :180.00   Max.   :140.0   Max.   :3.000  
```

γ. Πλέον είμαστε σε θέση να βρούμε ποιος τρόπος είναι κατάλληλος για κάθε μεταβλητή.

- **Height.m:**Μέση Τιμή + Τυπική Απόκλιση. Η μέση τιμή είναι πολύ κοντά στη διάμεση τιμή, η τυπική απόκλιση είναι πολύ μικρή και η κατανομή είναι ομοιόμορφη.
- **Weight.kg:** Σύνοψη των 5 αριθμών. Η τυπική απόκλιση είναι τεράστια και περνάει με μεγάλη διαφορά τη μέση τιμή. Επίσης, η κατανομή δεν είναι ομοιόμορφη.
- **HP,Defense:** Σύνοψη των 5 αριθμών γιατί μέσω του boxplot βλέπουμε καλύτερα την κατανομή και είμαστε σίγουροι για τα διάφορα outliers.
- **Attack,Speed:** Μέση Τιμή + Τυπική Απόκλιση γιατί δεν υπάρχουν ατυπικές τιμές.
- **Evolutions:** Μέση Τιμή + Τυπική Απόκλιση. Η μέση τιμή είναι πολύ κοντά στη διάμεση τιμή, η τυπική απόκλιση είναι πολύ μικρή και η κατανομή είναι ομοιόμορφη.

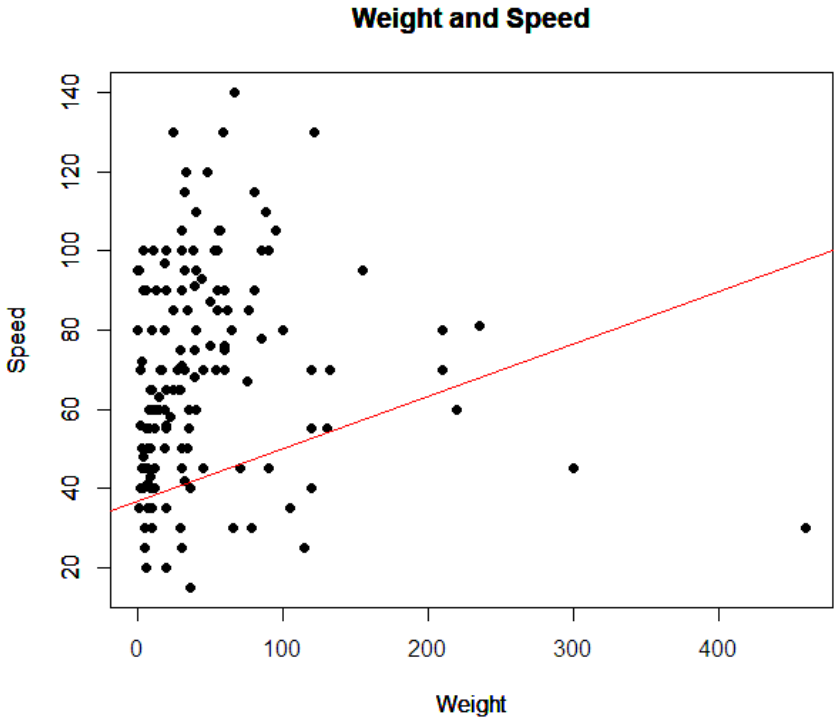
e. Θα διερευνήσουμε τη σχέση μεταξύ των μεταβλητών **Weight.kg.** και **Speed**. Με πρώτη σκέψη περιμένουμε πως όσο πιο βαρύ είναι ένα Pokémon τόσο πιο αργό μπορεί να είναι. Οπότε, αποθηκεύουμε τις μεταβλητές σε προσωρινές μεταβλητές x,y και δημιουργούμε το scatterplot:

```
> x <- MyTable$Weight.kg.
> y <- MyTable$Speed
> plot(x,y,main=Weight.kg. and Speed”,xlab=”Weight.kg.”,ylab=”Speed”,pch=16)
```

Έπειτα εφαρμόζουμε τη γραμμική παλινδρόμηση στο scatterplot:

```
> abline(lm(x ~ y),col="red")
```

Το αποτέλεσμα που παίρνουμε είναι:



Χρησιμοποιώντας τη συνάρτηση cor(x,y) παίρνουμε :

```
> cor(x,y)
[1] 0.0593211
```

Άρα, αφού ο συντελεστής συσχέτισης ισούται με αυτό τον αριθμό ο οποίος είναι πάρα πολύ κοντά στο 0, επιβεβαιώνουμε ότι η δύναμη της σχέσης είναι ασθενής και η συσχέτιση των δύο αυτών μεταβλητών είναι **σχεδόν ανύπαρκτη**. Είναι πλέον φανερό κιόλας, γιατί παρατηρούμε στο σχήμα πως υπάρχουν Pokémon ίδιου βάρους που η ταχύτητα τους είναι τελείως διαφορετική (από πολύ χαμηλή μέχρι και πολύ ψηλή. Οι περιπτώσεις που ικανοποιούν την αρχική μας υπόθεση είναι πολύ λίγες, όπως για παράδειγμα το Snorlax (η τέρμα δεξιά κουκίδα) που είναι το πιο βαρύ από όλα και έχει αρκετά χαμηλή ταχύτητα (αλλά ακόμα και έτσι ξεπερνάει την ταχύτητα κάποιων ελαφρύτερων Pokémon).

3ο Ερώτημα

Αποφασίσαμε να χρησιμοποιήσουμε τις ποσοτικές μεταβλητές **prob** και **math** που δείχνουν τις βαθμολογίες στα μαθήματα Πιθανότητες και Μαθηματικά Ι αντίστοιχα. Αρχικά, βάζουμε τα δεδομένα σε data frames βγάζοντας τα NA στοιχεία ως εξής :

```
> data <- read.table("survey_data_2020.txt",header = TRUE, sep=\t')
> mathWithoutNA <- na.omit(data["math"])
> probWithoutNA <- na.omit(data["prob"])
```

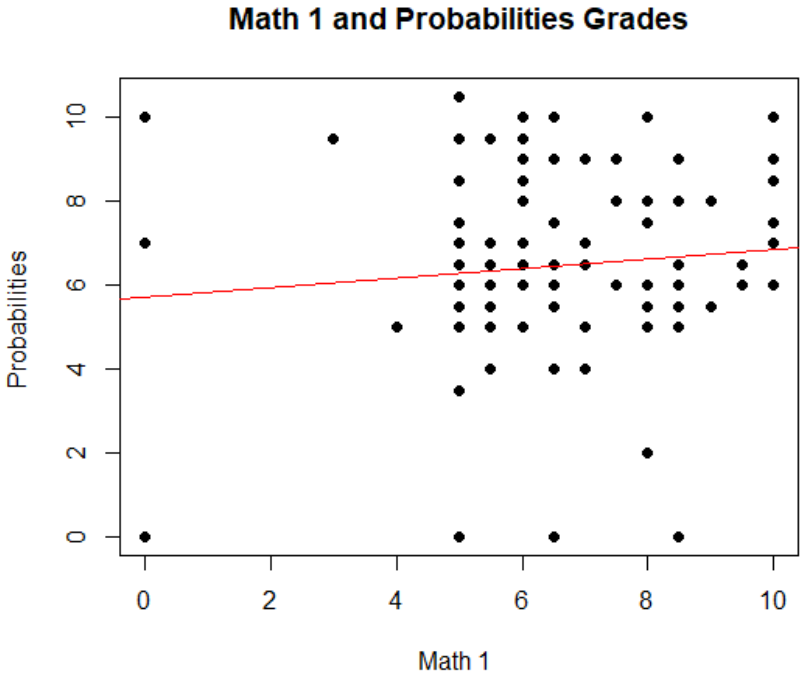
Αμέσως μετά μετατρέπουμε τα data frames σε vectors και δημιουργούμε το scatterplot :

```
> mathData <- mathWithoutNA$mathData
> probData <- probWithoutNA$probData
> plot(mathData,probData,main = "Math 1 and Probabilities Grades", xlab = "Math 1",
 ylab = "Probabilities", pch=16)
```

Εφαρμόζουμε και γραμμική παλινδρόμηση στο scatterplot :

```
> abline(lm(probData ~ mathData), col = "red")
```

Τέλος , έχουμε το εξής αποτέλεσμα :



a. Αρχικά , παρατηρούμε ότι η μορφή της σχέσης είναι γραμμική με πολύ μεγάλο πλήθος ατυπικών τιμών το οποίο καθιστά τη συγκεκριμένη σχέση ασθενή παρόλο που η γραμμή είναι αύξουσα. Άρα καταλήγουμε στα εξής :

Μορφή : Γραμμική / **Κατεύθυνση :** Αύξουσα /**Δύναμη** : Ασθενής

b. Χρησιμοποιώντας τη συνάρτηση cor(x,y) παίρνουμε :

```
>cor(mathData,probData)
[1] 0.1068717
```

Άρα, αφού ο συντελεστής συσχέτισης ισούται με αυτό το ποσό το οποίο είναι πάρα πολύ κοντά στο 0, επιβεβαιώνουμε ότι η δύναμη της σχέσης είναι ασθενής και η συσχέτιση των δύο αυτών μεταβλητών είναι **σχεδόν ανύπαρκτη**.