1η Άσκηση Στατιστική

ΓΕΩΡΓΙΟΥ ΑΛΕΞΙΟΣ ΛΑΖΑΡΟΣ 3180027

ΒΑΣΙΛΕΙΟΣ-ΕΚΤΩΡ ΚΩΤΣΗΣ-ΠΑΝΑΚΑΚΗΣ 3180094

1)

a. Χειρόγραφο (στο τέλος)

b. Για όλα τα dataframes την πιο ακριβή περιγραφή την κάνουν οι 5 αριθμοί

($min \leq\ Q1 \leq m \leq\ Q3 \leq\ \max$). Ωστόσο στο πρώτο dataframe, επειδή τα δεδομένα είναι συμμετρικά ως προς τη μέση τιμή και ακολουθούν την κανονική κατανομή μπορούμε να προσδιορίσουμε εξίσου καλά (και συνοπτικά) τα δεδομένα με μέση τιμή και την τυπική απόκλιση (m, s).

c. Βρισκω τη τυπική απόκλιση s για κάθε dataframe (Υπολογίστηκε μέσω της συνάρτησης sd() στην R)

Με βάση τον πίνακα τυποποιημένης κανονικής κατανομής $\Phi(z)\ =\ P(Z\ <=\ z)$, βρήκαμε ότι τα $Q1, Q3$ είναι περίπου $m\ \pm 0.67s$, δηλαδη το 25% αριστερα είναι στο $z\ =\ -0.67$ στην $N(0,1)$. Σχήμα

Υπολογίζουμε τα $Q1 = m - 0.67s,\ Q3\ =\ m\ +\ 0.67s$, οπού m η διάμεσος των δεδομένων.

Συγκρίνουμε τα πραγματικά $Q1, Q3$ με τα προσεγγιστικά $Q1', Q3'$.

Dataframe1) $s\ =\ 1.42, m\ =\ 32.55$, άρα

$$Q1'\ =\ 32.55 - 0.67\ *\ 1.42\ =\ 31.5986,$$

$$Q3'\ =\ 32.55\ +\ 0.67\ *\ 1.42\ =\ 33.5014$$

Τα οποία προσεγγίζουν πολύ κοντά τα πραγματικά $Q1 = 31.6$, $Q3 = 33.5$ και άρα η προσέγγιση της κατανομής των δεδομένων από μια καμπύλη πυκνότητας της Κανονικής κατανομής είναι ακριβής.

Dataframe2) $s\ =\ 3.06$, $m\ =\ 1.3$, άρα

$$Q1'\ =\ 1.3 - 0.67\ *\ 3.06 =\ -0.75,$$

$$Q3'\ =\ 1.3\ +\ 0.67\ *\ 3.06\ =\ 3.35$$

Τα οποία δεν προσεγγίζουν τα πραγματικά $Q1 = 0.5$, $Q3 = 3.7$ και άρα η προσέγγιση της κατανομής των δεδομένων από μια καμπύλη πυκνότητας της Κανονικής κατανομής είναι αποκλίνουσα. Η απόκλιση είναι πολύ μεγάλη για τη κλίμακα των δεδομένων μας (ή με βάση το IQR).

Dataframe3) $s\ =\ 28.26, m\ =\ 39.5$, άρα

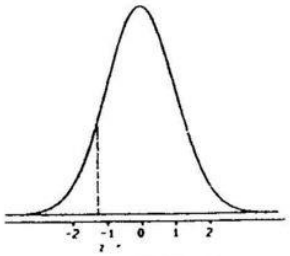
$$Q1'\ =\ 39.5 - 0.67\ *\ 28.26\ =\ 20.56,$$

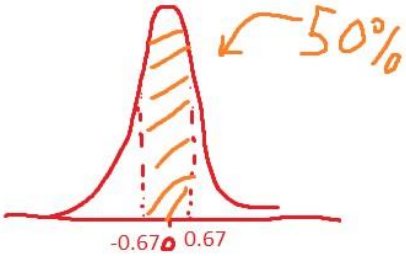
$$Q3'\ =\ 39.5\ +\ 0.67\ *\ 28.26\ =\ 58.43$$

Τα οποία προσεγγίζουν αρκετά κοντά τα πραγματικά $Q1 = 17.5$, $Q3 = 59$ και άρα η προσέγγιση της κατανομής των δεδομένων από μια καμπύλη πυκνότητας της Κανονικής κατανομής είναι μέτρια ακριβής. Η απόκλιση είναι μικρή για τη κλίμακα των δεδομένων μας (ή με βάση το IQR).

Πίνακας 2

Τυποποιημένη Κανονική Κατανομή





$\Phi(z) = P(Z \leq z)$

Τα στοιχεία του πίνακα εκφράζουν τις πιθανότητες $\Phi(z)=P(Z \leq z)$ που παριστάνονται από το εμβαδόν κάτω από την καμπύλη της τυποποιημένης κανονικής κατανομής αριστερά από το z.

	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-3.9	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
-3.8	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
-3.7	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
-3.6	0.0002	0.0002	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
-3.5	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002
-3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
-3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
-3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
-3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
-3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
-2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
-2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
-2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
-2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
-2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
-2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
-2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
-2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
-2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
-2.0	0.0227	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
-1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
-1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
-1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
-1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
-1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
-1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
-1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
-1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
-1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
-0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
-0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1921	0.1894	0.1867
-0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
-0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
-0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
-0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
-0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
-0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
-0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
-0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641

(Συνεχίζεται)

2)

a. Τα επιλεγμένα δεδομένα προέρχονται από την ιστοσελίδα vgchartz.com και αναφέρονται σε πωλήσεις ηλεκτρονικών παιχνιδιών από το 1987 μέχρι το 2017 (παγκόσμιες και ανά ήπειρο). Στα δεδομένα περιέχονται και χαρακτηριστικά για το κάθε παιχνίδι, όπως η κονσόλα στην οποία ανήκει, ο εκδότης κ.α.

Full link: https://data.world/julienf/video-games-global-sales-in-volume-1983-2017/workspace/file?filename=vgsalesGlobale.csv

b.

Από όλες τις μεταβλητές που συναντάμε αποφασίσουμε να χρησιμοποιήσουμε τις 4 παρακάτω για να αναλύσουμε:

Κατηγορικές μεταβλητές είναι οι ακόλουθες: **Year, Genre**

Ποσοτικές μεταβλητές είναι οι ακόλουθες: **Global_Sales, NA_SALES**

Θα κάνουμε μία συνοπτική εξήγηση των μεταβλητών που συναντάμε στο .csv μας

Year: Η χρονιά που κυκλοφόρησε το παιχνίδι. Παίρνει τιμες από 1987 μέχρι 2017.

Genre: Το είδος του παιχνιδιού (π.χ Δράσης, Μυστηρίου κτλ)

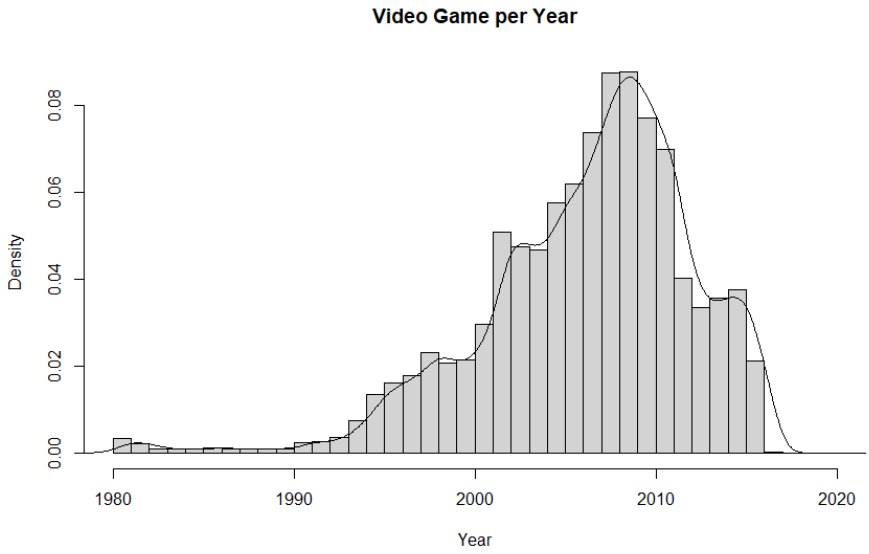
Global_Sales: Οι παγκόσμιες πωλήσεις του παιχνιδιού (Στα εκατομμύρια)

NA_Sales: Οι πωλήσεις του παιχνιδιού στην Βόρεια Αμερική (Στα εκατομμύρια)

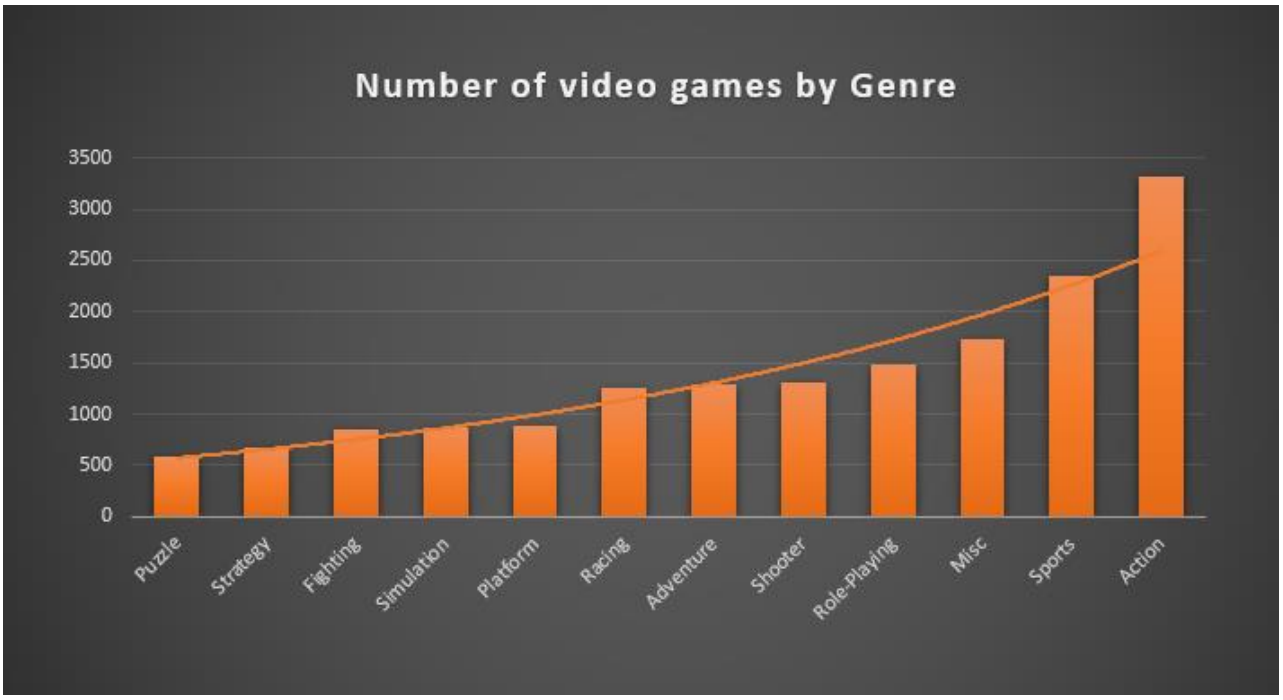
c. Τα παρακάτω διαγράμματα των επιλεγμένων μεταβλητών είναι χωρίς τα outliers τους, που βρέθηκαν με βάση τον κανόνα x1.5 IQR

c.

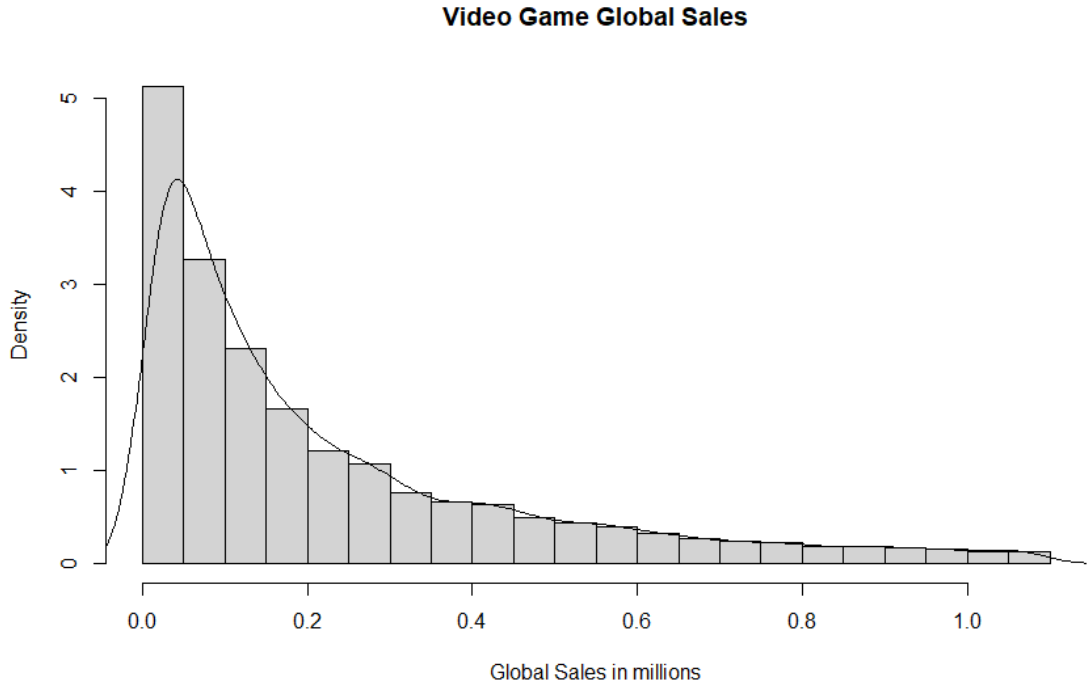
Year: Η κατανομή φαίνεται να προσεγγίζει την κανονική (οι περισσότερες τιμές φαινεται να είναι συσσωρευμένες γύρω από τις χρονιες 1997,1998). Το γεγονός ότι μετα αυτές τις 2 χρονιες υπάρχει πτώση μπορεί να οφείλεται σε ελλειπή δεδομένα.



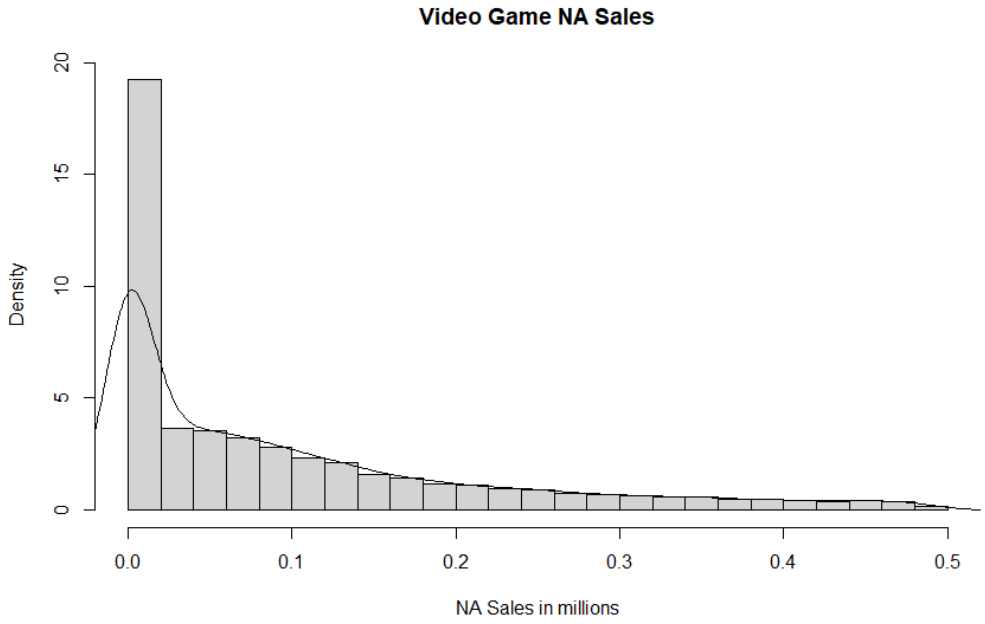
Genre: Η κατανομή δεν είναι ομοιόμορφη μεταξύ των τιμων αυτης της κατηγορικής μεταβλητής. Τα περισσότερα ηλεκτρονικα παιχνίδια φαίνεται να είναι δράσης και αθλητικά.Τα δεδομένα είναι ταξινομημένα σε αυξουσα σειρα



Global_Sales: Παρατηρούμε ότι τα παιχνίδια που εχουν μικρές πωλήσεις είναι πολύ λιγοτερα από αυτά που έχουν πολλές, όπως μπορουμε να δουμε από την παρακάτω κατανομή που περιγράφεται από μια φθίνουσα καμπύλη.



NA_Sales: Ακριβώς η ιδια παρατήρηση με την Global_Sales, απλώς για την Βόρεια Αμερική



---

d.

Global_Sales:

Mean = 0.5374, s = 1.555028

Min = 0.01 , Q1 =0.06, m = 0.17, Q3 = 0.47, Max = 82.74

Επειδή η κατανομή είναι κανονική, χρησιμεύει πιο πολύ η μέση τιμή με την τυπική απόκλιση.

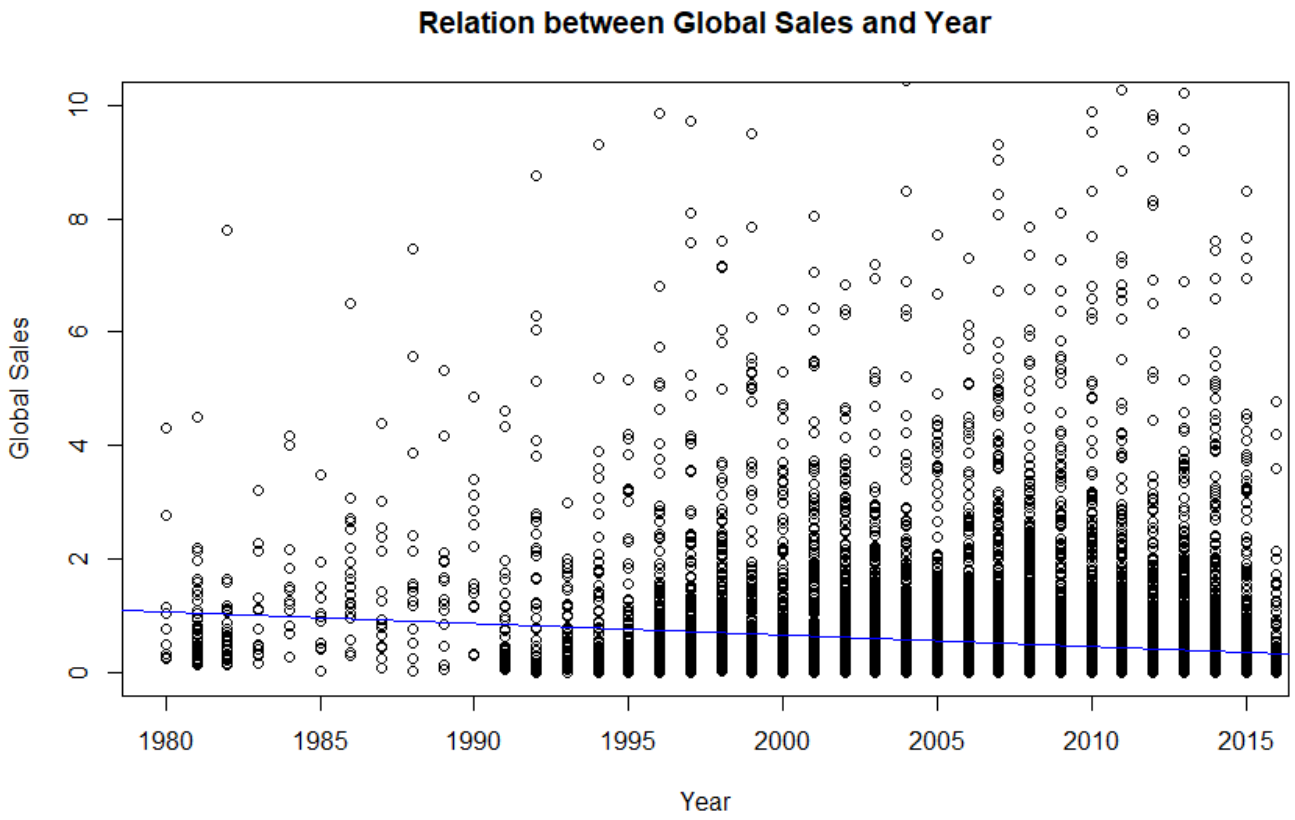
NA_Sales:

Mean = 0.2647, s = 0.816683

Min = 0, Q1 = 0, m = 0.8, Q3 = 0.24, Max = 41

Επειδή η κατανομή είναι ομοιόμορφη, χρησιμεύει πιο πολύ η σύνοψη των πέντε αριθμών.

e. Όπως βλέπουμε και από το παρακάτω διαγραμμα συσχέτισης (Μεταβλητή αιτίου = Year, Μεταβλητή αποτελέσματος= Global_Sales), η μεταβλητή Year είναι ασθενώς φθίνουσα συσχετισμένη με την μεταβλητή Global_Sales. Αυτό μπορούμε να το δουμε και από την μπλε ευθεία γραμμή παλινδρόμισης ελαχίστων τετραγώνων. Συνεπώς, όσο πιο πρόσφατη η χρονολογία, τόσο πιο πιθανό είναι ένα παιχνίδι να πωλήσει λίγο. Αυτό όμως δε σημαίνει ότι η σχεση των δυο μεταβλητών είναι αιτιατή, καθώς υπάρχουν κρυφοί παράγοντες. Δύο από αυτούς είναι η προσβασιμότητα στην δημιουργία παιχνιδιών και το συνολικό πλήθος των παιχνιδιών, καθώς δεν μπορούν όλα τα προϊόντα σε μια αγορά να διαμοιράζονται εξίσου τις πωλήσεις των καταναλωτών (Δεν είναι βιώσιμο για τον καταναλωτή να αγοράσει όλα τα παιχνίδια). Επίσης, κοιτώντας το διάγραμμα μπορούμε να συμπαιράνουμε πως τα πιο πρόσφατα επιτυχημένα παιχνίδια εχουν κατά μέσο όρο περισσότερες πωλήσεις από ότι έχουν τα πιο παλία επιτυχημένα παιχνίδια, πράγμα που οφείλεται στην αυξημένη δημοτικότητα που έχουν αποκτήσει και όχι άμεσα στην χρονιά

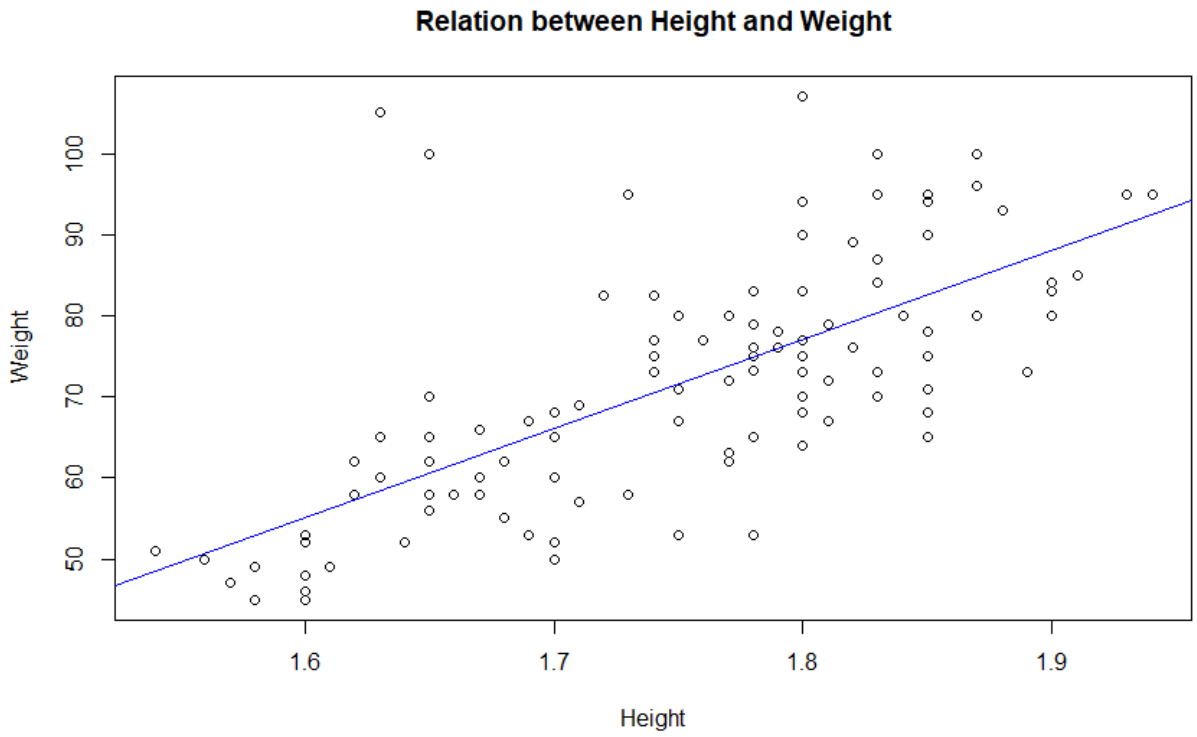


Έχει γίνει εστίαση στο διάγραμμα.

3)

a. Κάνουμε import τα survey_data_2020 από το txt file. Εκτελώντας το παρακάτω κώδικα στην R, φτιάχνουμε το scatterplot.

```
> data <- survey_data_2020
> data <- data[c('height', 'weight')] #Keeping only these columns in our
dataframe
> data <- data[rowSums(is.na(data)) != ncol(data),] #Removing NAs
> data <- data[-c(7, 22,94),] #Removing outliers
> plot(data$height, data$weight, ylab = "Weight", xlab = "Height", main =
"Relation between Height and Weight")
```



Η μορφή του scatterplot είναι γραμμική καθώς φαίνεται στο scatterplot ότι τα στοιχεία δημιουργούν γραμμική συνάρτηση και η κατεύθυνση είναι αύξουσα εφόσον για τα περισσότερα στοιχεία, όσο το ύψος αυξάνεται, τόσο αυξάνεται και το βάρος. Τέλος η δύναμη είναι μέτρια ισχυρή, εφόσον η πυκνότητα της συσσώρευσης των σημείων για ορισμένα διαστήματα τιμών του ύψους είναι αρκετά μεγάλη, αλλά όχι σε βαθμό που να καθιστά την δύναμη ισχυρή.

b. Μετά τον υπολογισμό του στην γλώσσα R, και μετά την αφαίρεση των outliers και άκυρων τιμών, ο συντελεστής συσχέτισης βρέθηκε ίσος με 0.69. Πράγμα που επιβεβαιώνει του προηγουμένους ισχυρισμούς μας. `> cor(data$height, data$weight)`

Η γραμμική παλινδρόμηση φαίνεται στο παραπάνω σχήμα (μπλε γραμμή).

```
#Linear Regression calculation and plotting
> weight <- data$weight
> height <- data$height
> model <- lm(weight~height)
> model

Call:
lm(formula = weight ~ height)

Coefficients:
       (Intercept)  height
            -120.5   109.8
```

```
> abline(model, col = 'blue')
```

1η Σειρά Ασκήσεων, Στατιστική

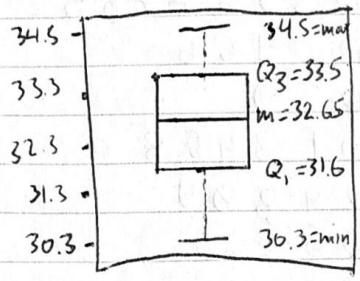
ΓΕΩΡΓΙΟΥ ΑΛΕΞΙΟΣ-ΛΑΖΑΡΟΣ 3180027
ΚΩΤΣΗΣ ΠΑΝΑΓΙΩΤΗΣ ΒΑΣΙΛΕΙΟΣ-ΕΚΤΩΡ 3180094

1.

α) STEMPLOT 1

30	3
31	01
32	167
33	46
34	25

BOXPLOT 1



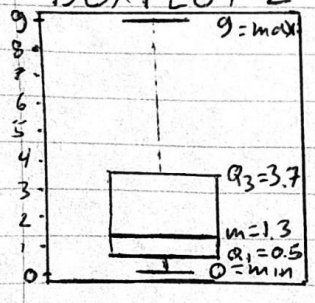
max = 34.5
Q_3 = 33.5
m = 32.65
Q_1 = 31.6
min = 30.3

STEMPLOT 2

0	0028
1	24
2	
3	2
4	2
5	
6	4
7	
8	
9	0

0-2	002824
3-5	22
6-8	4
9	0

BOXPLOT 2



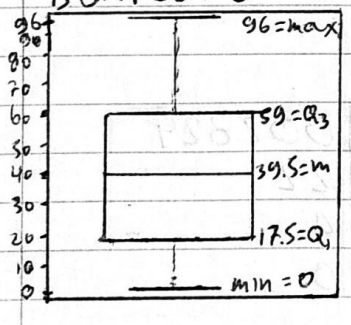
max = 9
Q_3 = 3.7
m = 1.3
Q_1 = 0.5
min = 0

STEMPLOT 3

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

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

BOXPLOT 3



max = 96
Q_3 = 59
m = 39.5
Q_1 = 17.5
min = 0