ΝΙΚΟΛΑΟΣ ΜΗΤΣΑΚΗΣ [ΠΛΗΡ]
Α.Μ. : 3210122

1η Σειρά Ασκήσεων – Στατιστική στην Πληροφορική, 2023-2024

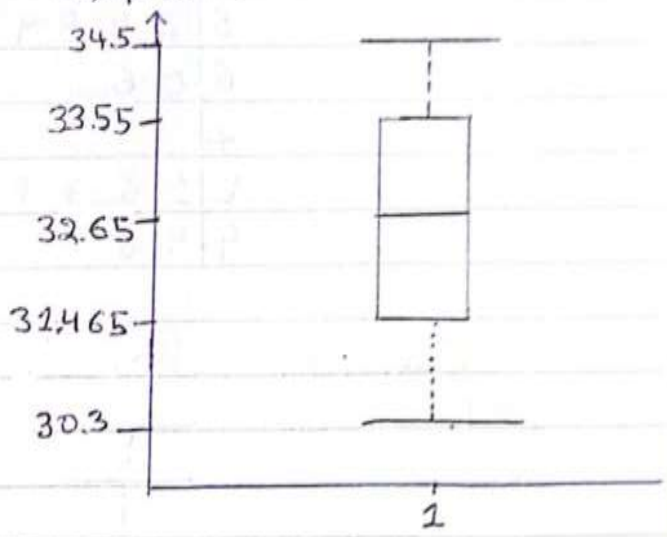
Άσκηση 1).

α). Δεδομένα Ι: 30.3, 31.0, 31.1, 32.1, 32.6, 32.7, 33.4, 33.6, 34.2, 34.5

Stemplot:

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

Boxplot:



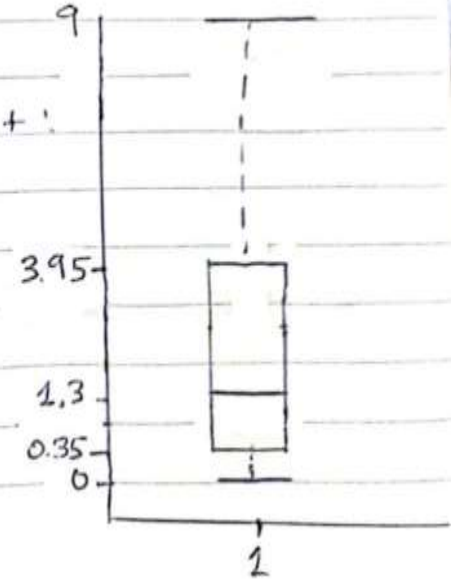
$Q_1 = 31.465$
$Q_2 = 32.65$
$Q_3 = 33.55$
min = 30.3
max = 34.5

Δεδομένα ΙΙ: 0.0, 0.0, 0.2, 0.8, 1.2, 1.4, 3.2, 4.2, 6.4, 9.0

Stemplot:

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

Boxplot:



$Q_1 = 0.35$ min = 0
$Q_2 = 1.3$ max = 9
$Q_3 = 3.95$

ΝΙΚΟΛΑΟΣ ΜΗΤΣΑΚΗΣ [ΠΛΗΡ]
Α.Μ. : 3210122

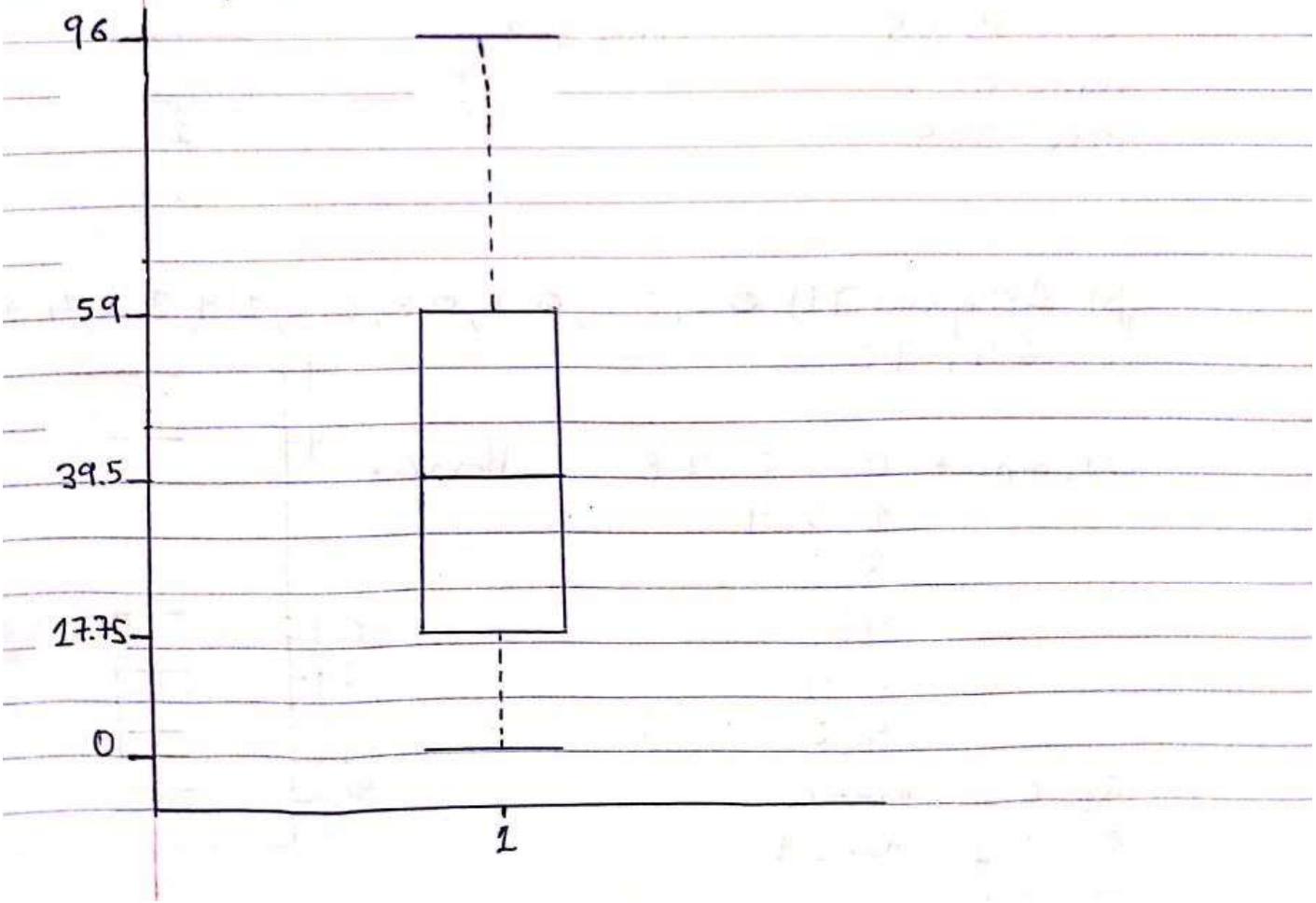
Δεδομένα III: 0, 1, 6, 8, 10, 13, 15, 16, 17, 17, 18, 18, 20, 20, 21, 25, 26, 30, 35, 41, 43, 44, 46, 48, 52, 54, 58, 59, 59, 60, 66, 81, 86, 87, 88, 89, 94, 96

Stemplot:

0	0 1 6 8
1	0 3 5 6 7 7 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

max = 96
min = 0
$Q_1 = 17.75$
$Q_2 = 39.5$
$Q_3 = 59$

Boxplot:



ΝΙΚΟΛΑΟΣ ΜΗΤΣΑΚΗΣ [ΠΛΗΡ]
Α.Μ. : 3210122

b).

Δεδομένα I (Data1):

Το εύρος των δεδομένων είναι μικρό (34.5 − 30.3 = 4.2), και η τυπική απόκλιση είναι επίσης χαμηλή (1.34). Σε αυτή την περίπτωση, η μέση τιμή και η τυπική απόκλιση είναι αρκετές για να περιγράψουν την κεντρική τάση και τη διασπορά για αυτό το σύνολο δεδομένων.

Δεδομένα II (Data2):

Τα δεδομένα έχουν μεγάλο εύρος (9.0 − 0.0 = 9.0) και υψηλή τυπική απόκλιση (2.90). Η σύνοψη των 5 αριθμών παρέχει περισσότερες πληροφορίες για το άπλωμα των δεδομένων και τα τεταρτημόρια, κάνοντάς την πιο περιγραφική για αυτό το σύνολο δεδομένων.

Δεδομένα III (Data3):

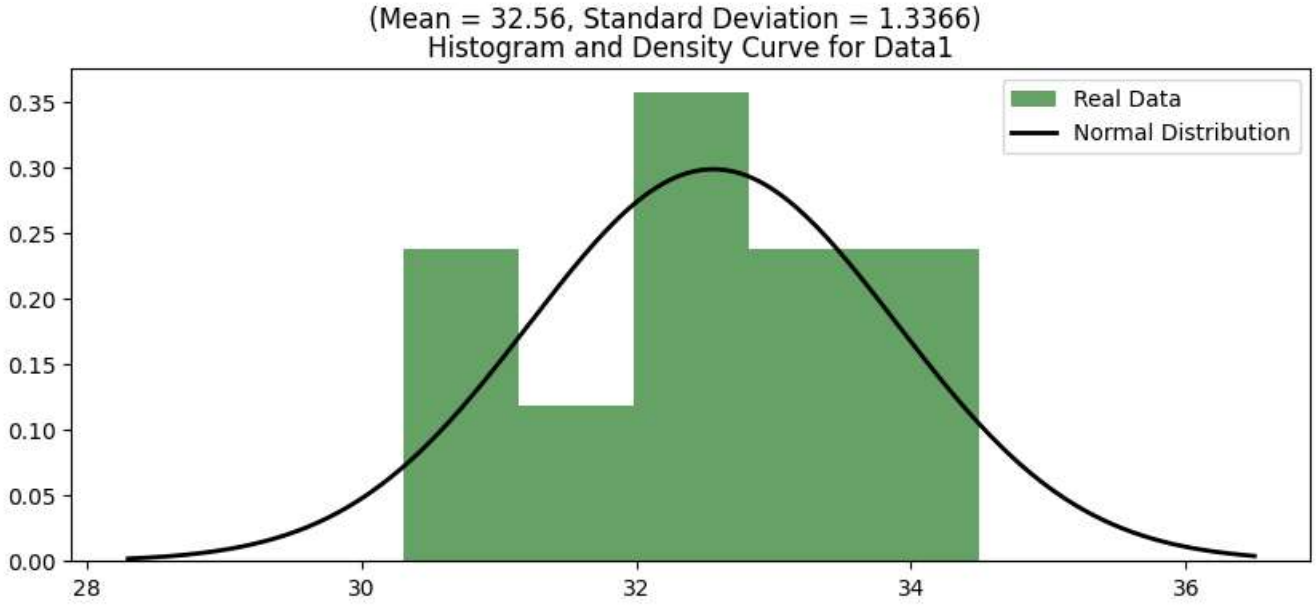
Το εύρος είναι αρκετά μεγάλο (96 − 0 = 96) και η τυπική απόκλιση είναι επίσης υψηλή (27.91). Δεδομένου του έντονου απλώματος που παρουσιάζουν τα δεδομένα, η σύνοψη των 5 αριθμών θα ήταν πιο περιγραφική για την κατανόηση της κατανομής του συγκεκριμένου συνόλου δεδομένων.

- Συνοπτικά, για τα Data1, η μέση τιμή και η τυπική απόκλιση είναι αρκετά πληροφοριακές, ενώ για τα Data2 και Data3, η σύνοψη των 5 αριθμών προσφέρει μία πιο περιγραφική ανάλυση.

c).

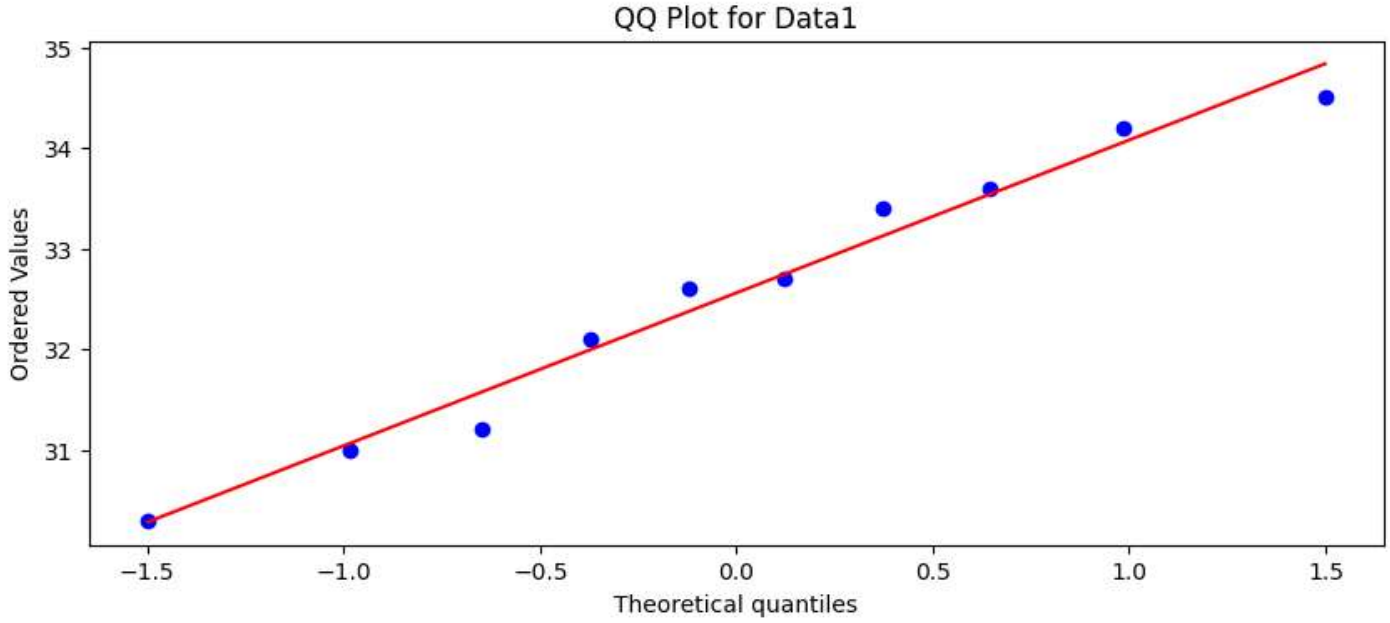
Δεδομένα I (Data1):

- Το ιστόγραμμα δείχνει ότι τα δεδομένα ταιριάζουν στενά με την καμπύλη της κανονικής κατανομής, υποδηλώνοντας καλή προσαρμογή.



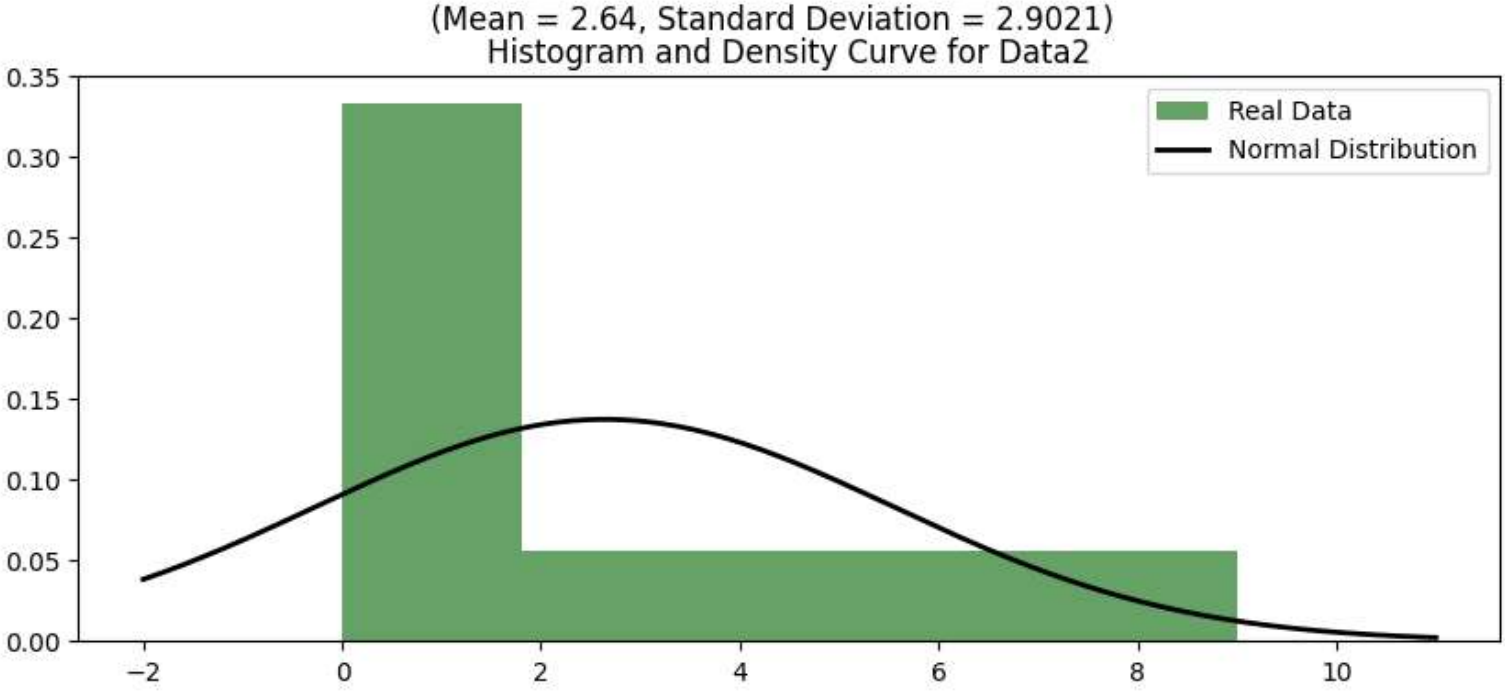
ΝΙΚΟΛΑΟΣ ΜΗΤΣΑΚΗΣ [ΠΛΗΡ]
Α.Μ. : 3210122

- Το Q-Q διάγραμμα δείχνει ότι τα σημεία βρίσκονται κοντά στην ευθεία, υποδεικνύοντας ότι τα ποσοστημόρια των δεδομένων είναι παρόμοια με αυτά που θα περιμέναμε από μια κανονική κατανομή.

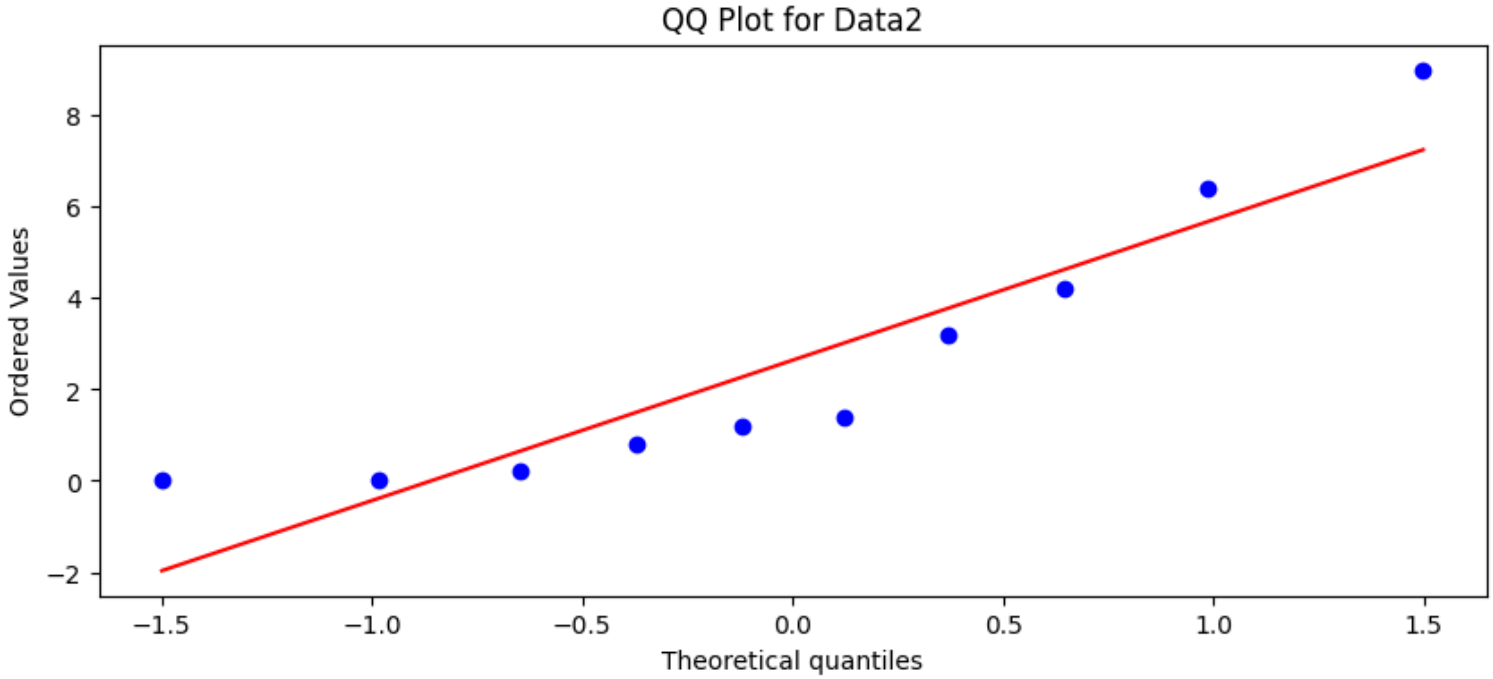


<u>Δεδομένα ΙΙ (Data2):</u>

- Το ιστόγραμμα δείχνει ότι η κατανομή του data2 είναι μη συμμετρική με πολλές τιμές κοντά στο μηδέν, κάτι που δεν είναι χαρακτηριστικό μιας κανονικής κατανομής.

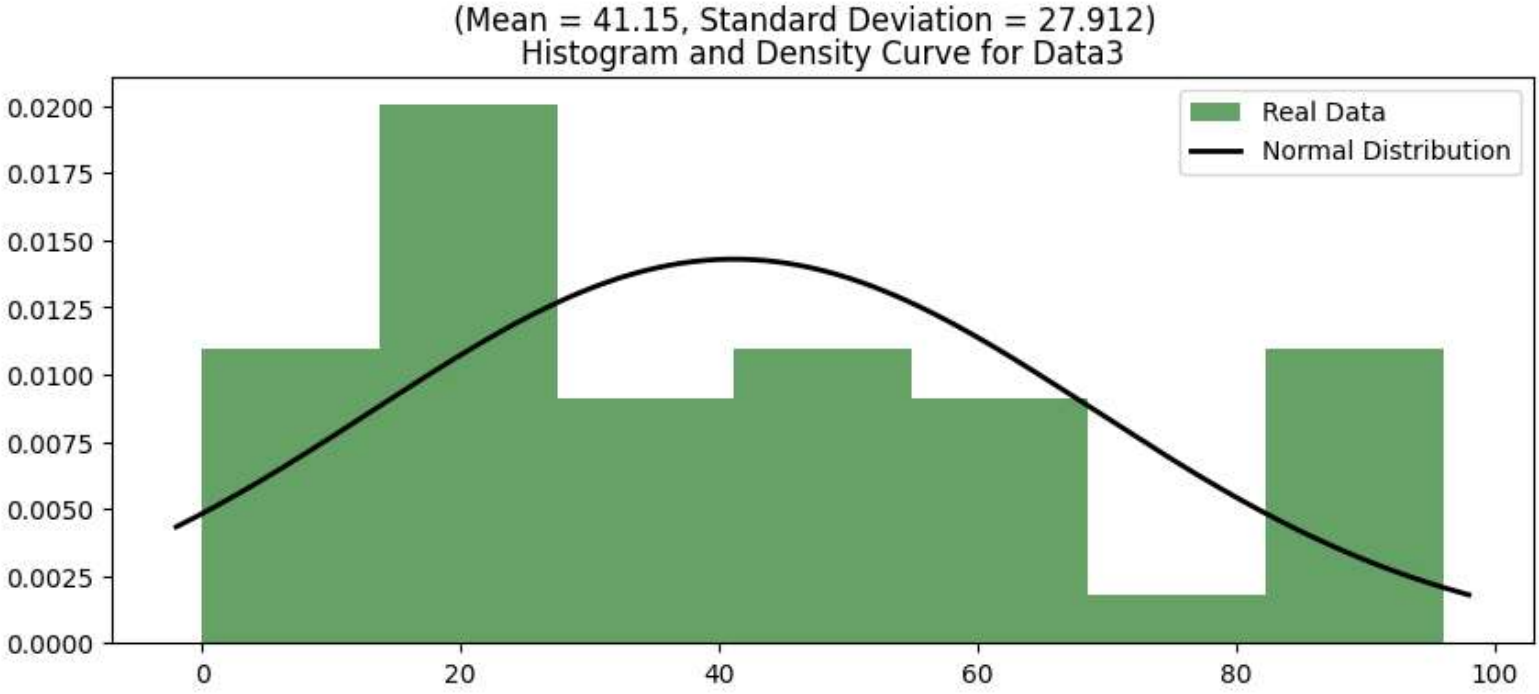


- Το Q-Q διάγραμμα δείχνει απόκλιση από την ευθεία κυρίως στα άκρα, ιδιαίτερα στο κατώτερο τμήμα, υποδηλώνοντας ότι η κατανομή δεν είναι κανονική.

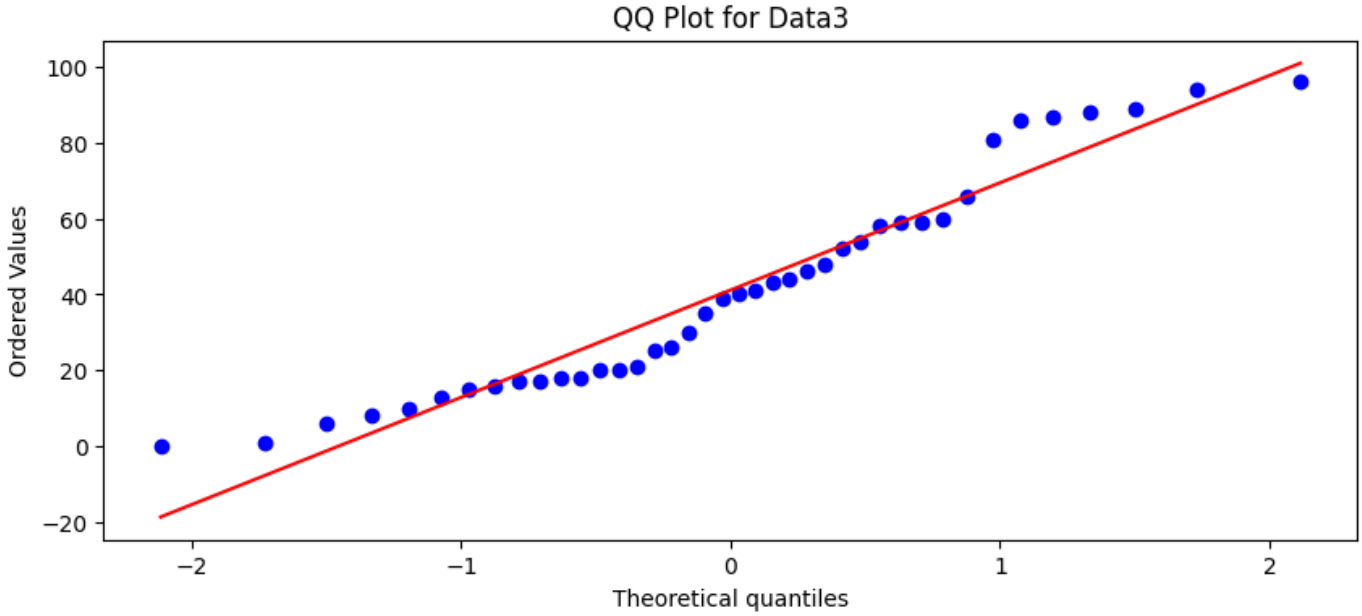


<u>Δεδομένα ΙΙΙ (Data3):</u>

- Το ιστόγραμμα αποκαλύπτει μια πιο ευρεία κατανομή με μια πιθανή ακραία τιμή στο υψηλότερο άκρο, που μπορεί να επηρεάσει το ταίριασμα με τη κανονική κατανομή.



- Το Q-Q διάγραμμα δείχνει μια καμπύλη, αποκλίνοντας από την ευθεία, ειδικά στα δύο άκρα, υποδηλώνοντας ότι το data3 δεν ακολουθεί στενά μια κανονική κατανομή.



- Συνοψίζοντας, το Data1 φαίνεται να προσεγγίζεται καλά από τη Κανονική Κατανομή, ενώ το Data2 και το Data3 εμφανίζουν σημαντικές αποκλίσεις, ιδιαίτερα στις ουρές τους. Αυτές οι αποκλίσεις δείχνουν ότι η Κανονική Κατανομή δεν αποτελεί ακριβής προσέγγιση για το Data2 και το Data3, και άλλες κατανομές μπορεί να είναι πιο κατάλληλες για αυτά τα σύνολα δεδομένων.

Άσκηση 2).

a) Πληροφορίες:
Πηγή: https://corgis-edu.github.io/corgis/csv/video_games/
Ημερομηνία έκδοσης: 21 May 2015
Ημερομηνία παραγωγής δεδομένων: Jan 2013
Εκδότης: University of Portsmouth
Πλήθος Εγγραφών:1212

- **Δείτε περισσότερες πληροφορίες πατώντας το παρακάτω σύνδεσμο: https://researchportal.port.ac.uk/en/publications/what-makes-a-blockbuster-video-game-an-empirical-analysis-of-us-s**
- **Για να κατεβάσετε το dataset πατήστε το παρακάτω σύνδεσμο: https://researchportal.port.ac.uk/files/2366486/Managerial_and_Decision_Economics_2013_Video_Games_Dataset.csv**

ΝΙΚΟΛΑΟΣ ΜΗΤΣΑΚΗΣ [ΠΛΗΡ]
Α.Μ. : 3210122

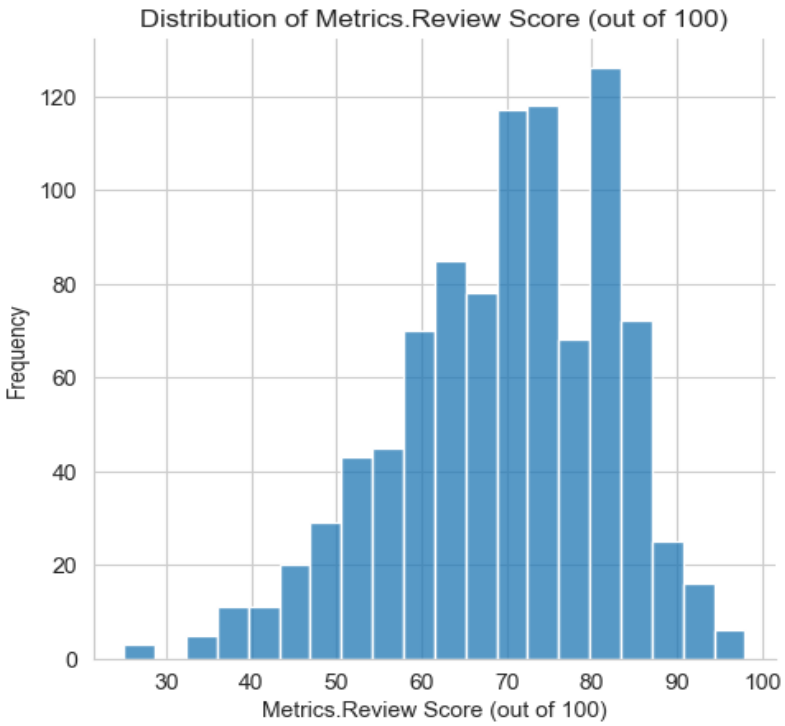
b)

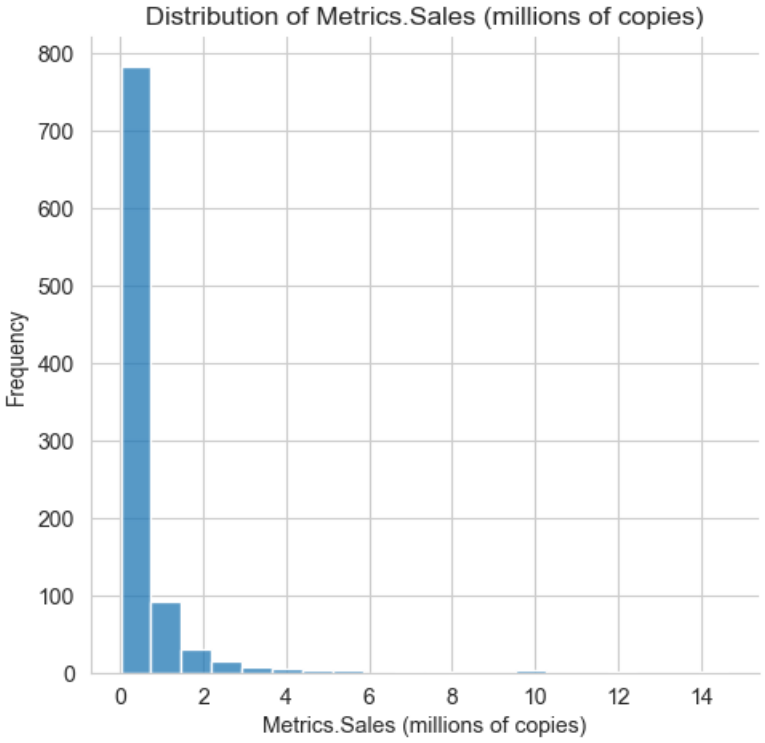
- **<u>Κατηγορικές Μεταβλητές:</u>**
  - Title, type-> object
    - Ο τίτλος του βιντεοπαιχνιδιού.
  - Features.Handheld?, type-> bool
    - Έχει κυκλοφορήσει το βιντεοπαιχνίδι σε hanheld κονσόλα (πχ. PSP);
  - Features.Multiplatform?, type-> bool
    - Έχει κυκλοφορήσει το βιντεοπαιχνίδι για πολλαπλές πλατφόρμες;
  - Features.Online?, type-> bool
    - Υποστηρίζει το βιντεοπαιχνίδι την σύνδεση με το διαδίκτυο;
  - Metadata.Genres, type-> object
    - Σε ποιο ή ποια είδη ανήκει το βιντεοπαιχνίδι.
  - Metadata.Licensed?, type-> bool
    - Έχει λάβει το βιντεοπαιχνίδι όλες τις απαραίτητες νομικές άδειες για να χρησιμοποιήσει συγκεκριμένη πνευματική ιδιοκτησία;
  - Metadata.Sequel?, type-> bool
    - Έχει κυκλοφορήσει κάποιο άλλο παιχνίδι στον ίδιο «κόσμο» με το συγκεκριμένο βιντεοπαιχνίδι;
  - Release.Console, type-> object
    - Για ποια ή για ποιες κονσόλες έχει κυκλοφορήσει το βιντεοπαιχνίδι.
  - Release.Re-release?, type-> bool
    - Έχει γίνει κάποιου είδους επανέκδοση στην αγορά για το συγκεκριμένο βιντεοπαιχνίδι;
  - Release.Rating, type-> object
    - Σε ποια από τις παρακάτω κατηγορίες ηλικίας ή/και περιεχομένου έχει καταταγεί το συγκεκριμένο βιντεοπαιχνίδι.
      - E (Everyone): Παιχνίδια κατάλληλα για όλους.
      - T (Teen): Κατάλληλα για έφηβους 13 ετών και πάνω.
      - M (Mature): Κατάλληλα για άτομα 17 ετών και πάνω.

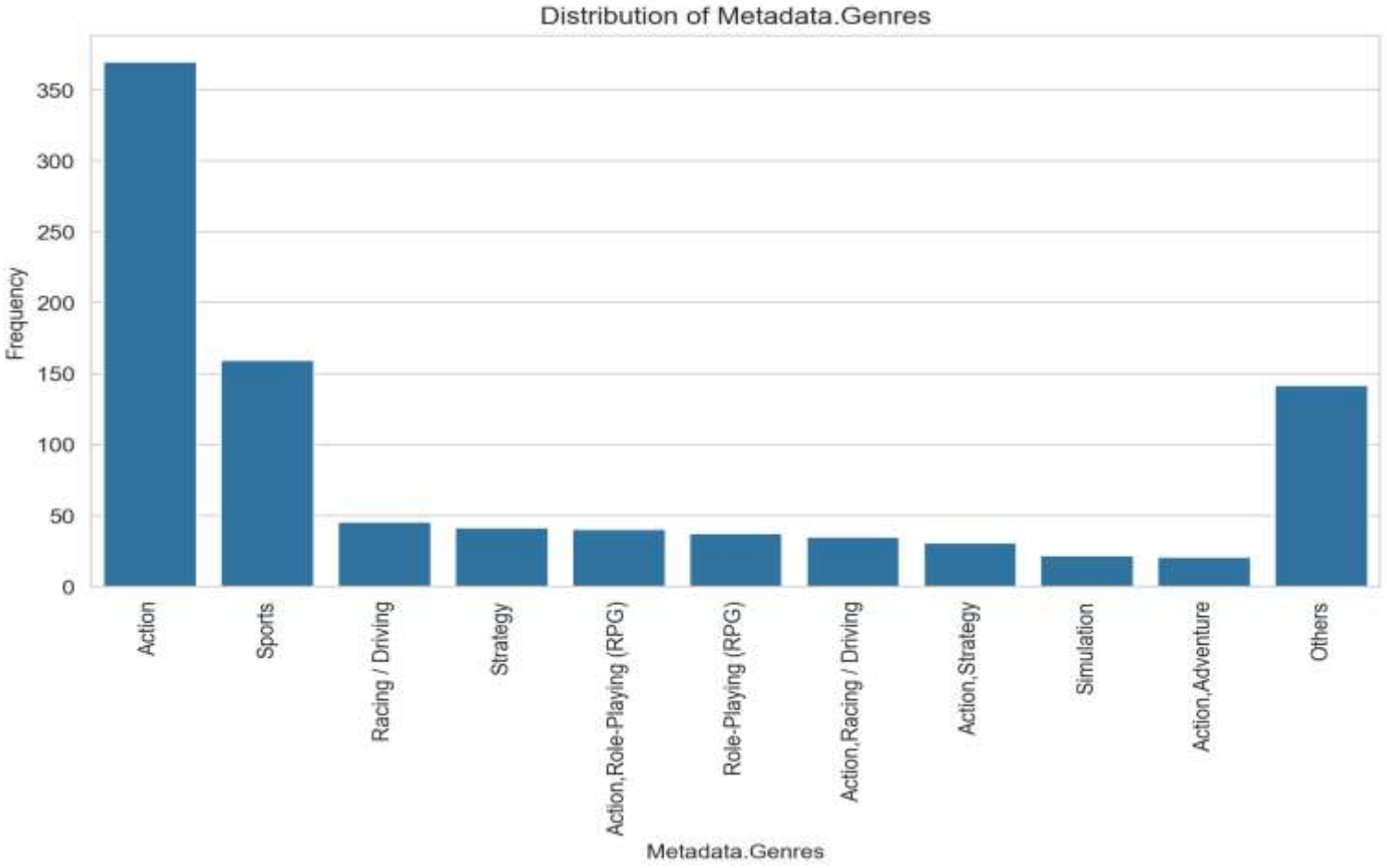
- **<u>Ποσοτικές Μεταβλητές:</u>**
  - Features.Max Players, type -> int64
    - Τα μέγιστο πλήθος (τοπικών) παικτών που υποστηρίζει το βιντεοπαιχνίδι.
  - Metrics.Review Score, type -> int64
    - Η βαθμολογία του βιντεοπαιχνιδιού.
  - Metrics.Sales, type -> float64
    - Οι πωλήσεις (σε εκατομμύρια) του βιντεοπαιχνιδιού.
  - Metrics.Used Price, type -> float64
    - Η τιμή του βιντεοπαιχνιδιού, μεταχειρισμένο.
  - Release.Year, type -> int64
    - Το έτος κυκλοφορίας του βιντεοπαιχνιδιού.

ΝΙΚΟΛΑΟΣ ΜΗΤΣΑΚΗΣ [ΠΛΗΡ]
Α.Μ. : 3210122

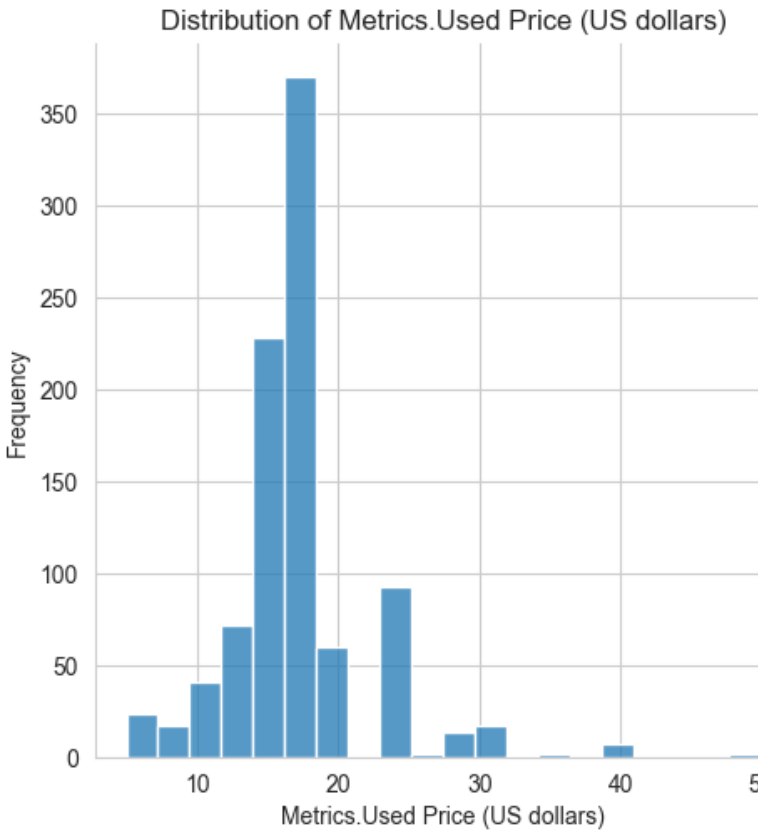
c) Κατανομές όλων των μεταβλητών του dataset(εκτός του Title):

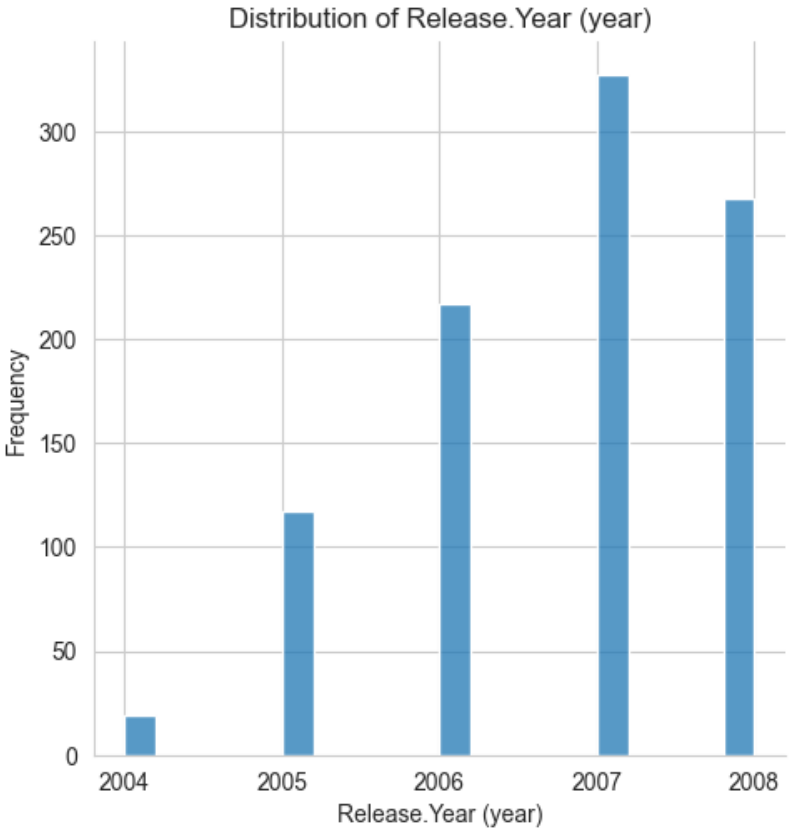


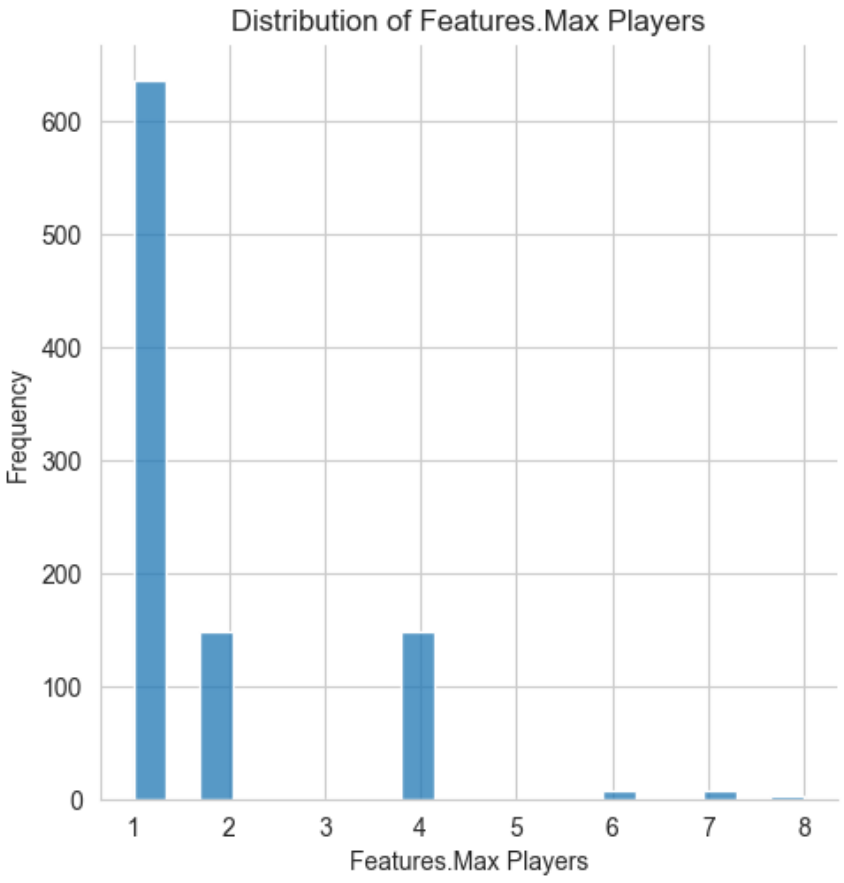


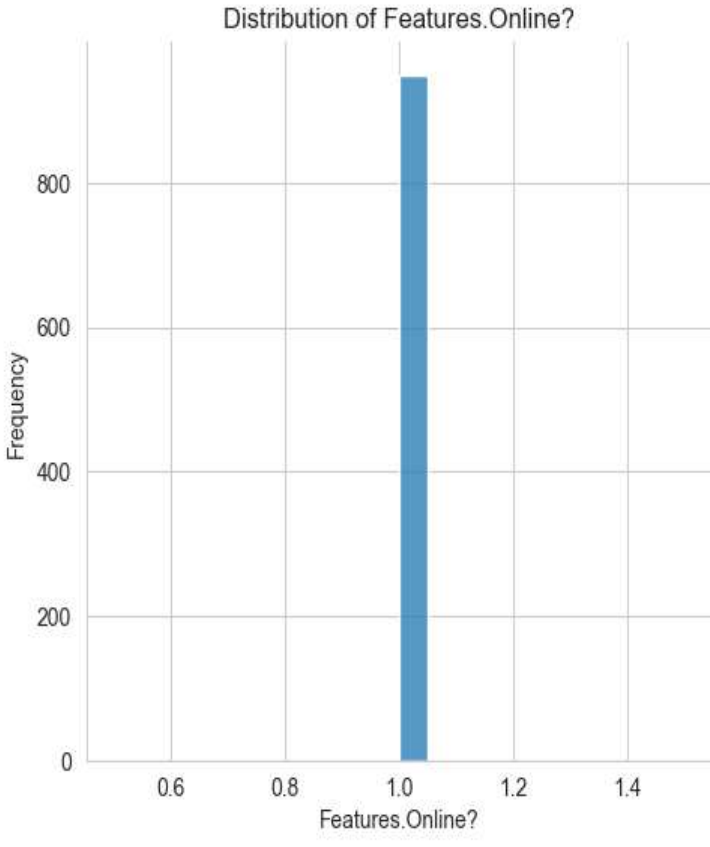


ΝΙΚΟΛΑΟΣ ΜΗΤΣΑΚΗΣ [ΠΛΗΡ]
Α.Μ. : 3210122

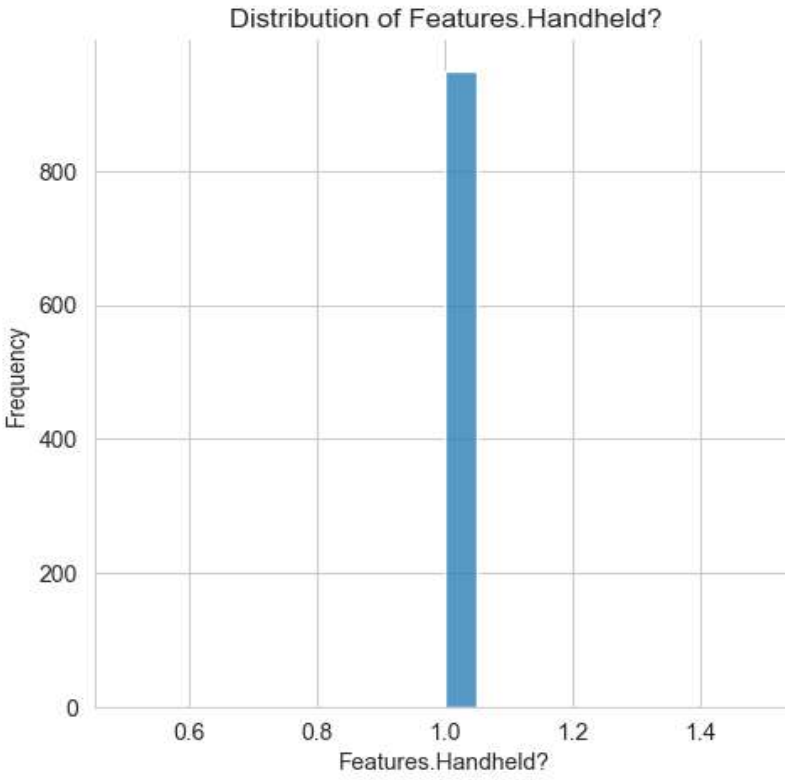


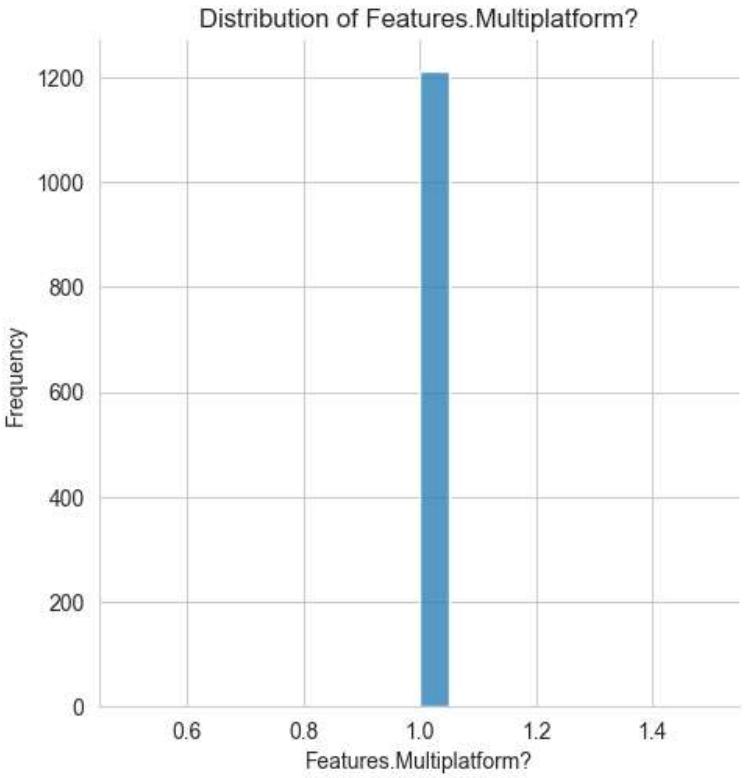


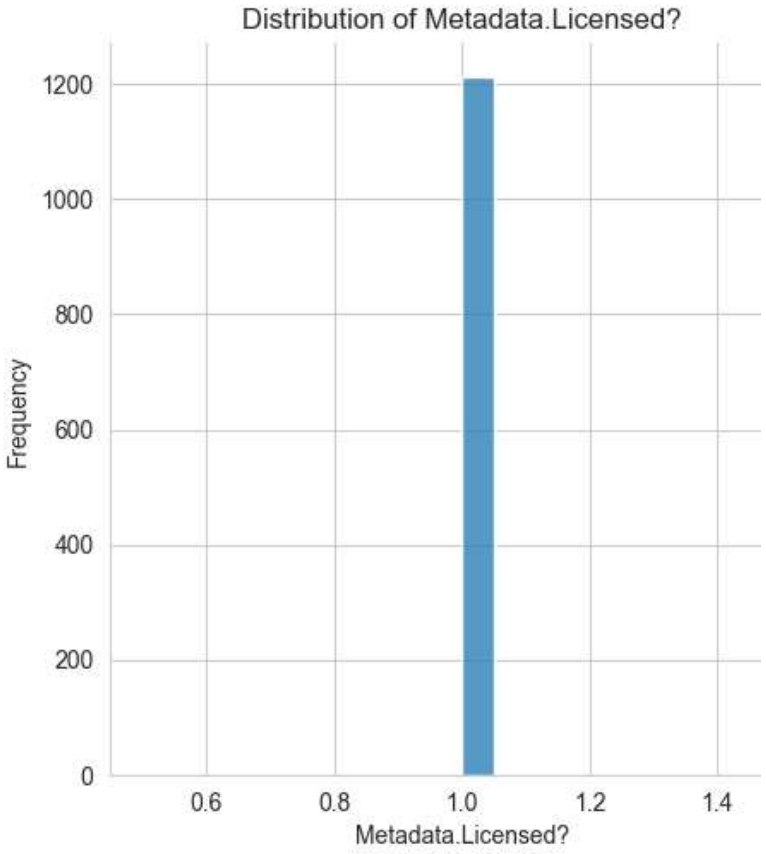


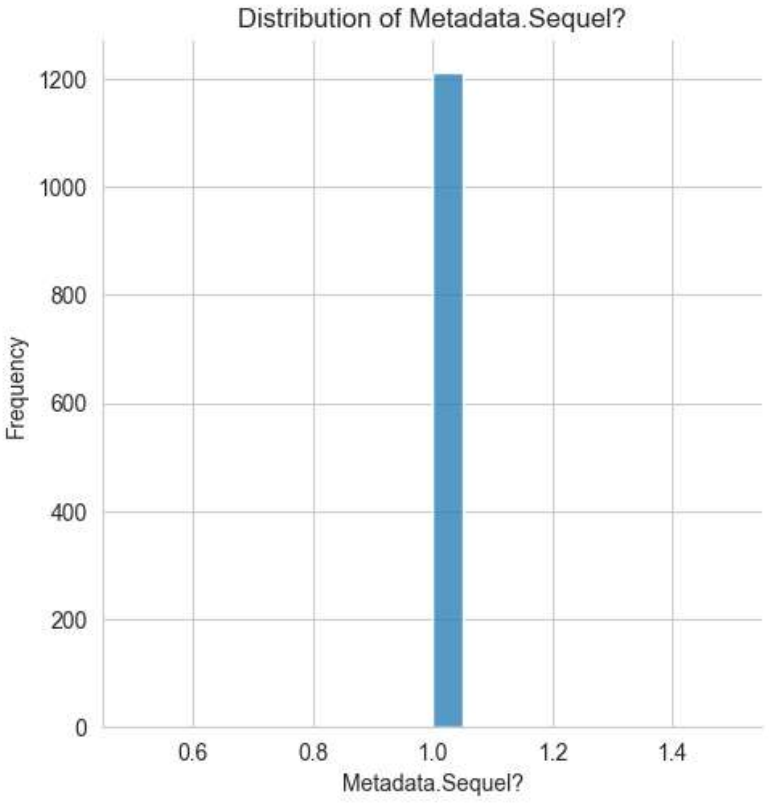


ΝΙΚΟΛΑΟΣ ΜΗΤΣΑΚΗΣ [ΠΛΗΡ]
Α.Μ. : 3210122

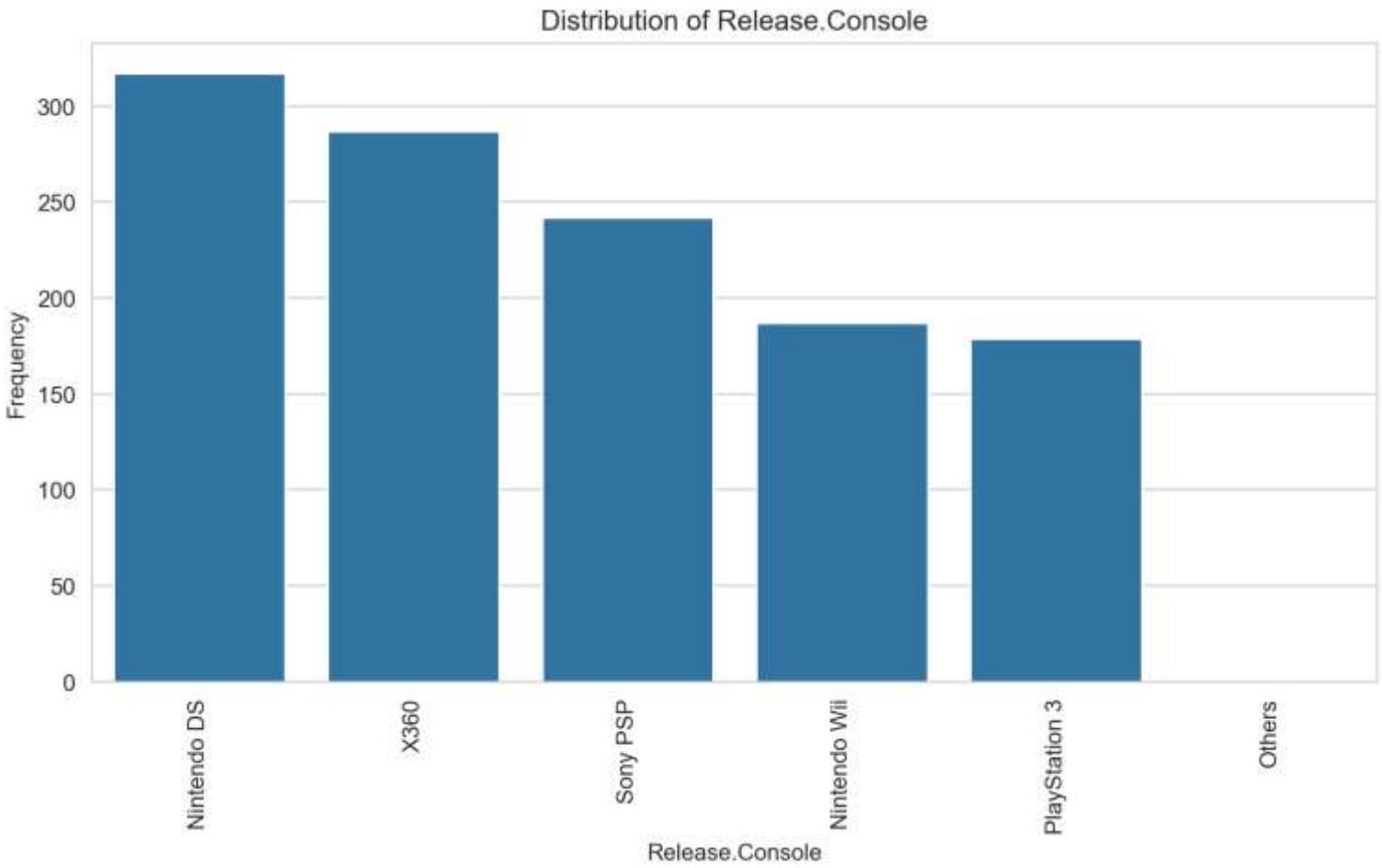


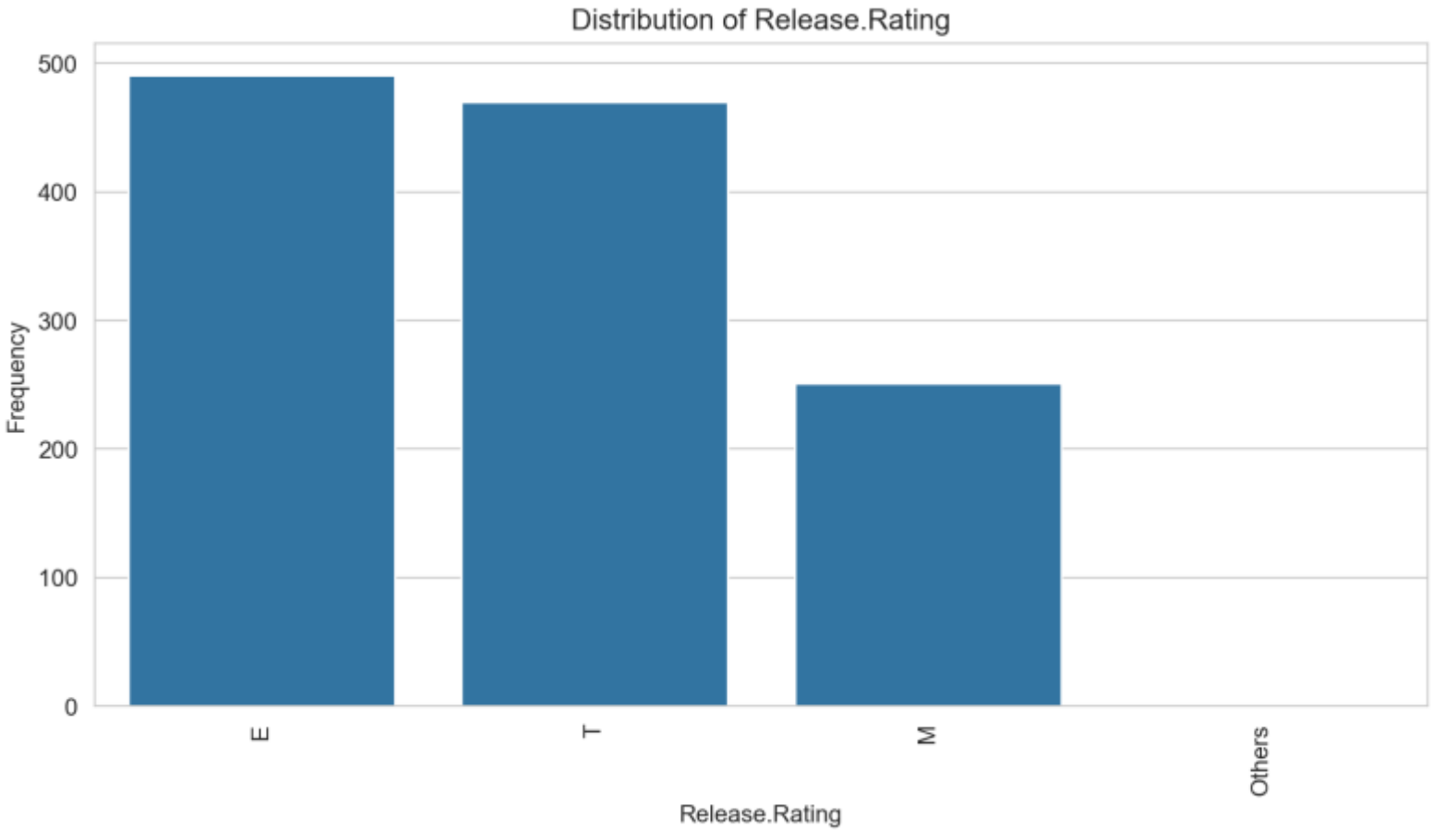






ΝΙΚΟΛΑΟΣ ΜΗΤΣΑΚΗΣ [ΠΛΗΡ]
Α.Μ. : 3210122





d) Η μέση τιμή, η τυπική απόκλιση και η σύνοψη των 5 αριθμών κάθε ποσοτικής μεταβλητής του dataset:

- -

- **Features.Max Players Average:**

Features.Max Players Average: 1.720464135021097

Features.Max Players Standard Deviation: 1.2639602685552929

- Features.Max Players 5-number summary:

| min = 1.0 |

| 25% = 1.0 |

| 50% = 1.0 |

| 75% = 2.0 |

| max = 8.0 |

Name: Features.Max Players, dtype: float64

- Για τη συγκεκριμένη μεταβλητή, η σύνοψη των 5 αριθμών είναι η καταλληλότερη για να υπογραμμίσει το γεγονός ότι τα δεδομένα είναι δεξιά-στρεβλωμένα, με τη μέγιστη τιμή (8) να είναι πολύ υψηλότερη από το 75% (2), δείχνοντας ότι οι περισσότερες τιμές είναι χαμηλές με κάποιες εξαιρέσεις ακραίων σημείων με υψηλές τιμές.

- -

- **Metrics.Review Score:**

Metrics.Review Score Average: 69.68143459915612

Metrics.Review Score Standard Deviation: 12.739357454423628

- Metrics.Review Score 5-number summary:

| min = 25.00 |

| 25% = 61.75 |

| 50% = 71.00 |

| 75% = 80.00 |

| max = 98.00 |

Name: Metrics.Review Score, dtype: float64

- Για τη συγκεκριμένη μεταβλητή, και οι δύο τρόποι κάνουν καλή δουλειά στο να περιγράψουν την μεταβλητή, με τη σύνοψη των 5 αριθμών να είναι ίσως πιο κατάλληλη διότι υποδηλώνει μία πιθανή αριστερή στρέβλωση στα δεδομένα, καθώς το μέσο (71) είναι υψηλότερο από τη μέση τιμή (69.68) και το εύρος είναι αρκετά μεγάλο (από 25 έως 98).

- -

- **Metrics.Sales:**

Metrics.Sales Average: 0.5620991561181434

Metrics.Sales Standard Deviation: 1.1736861142138202

- Metrics.Sales 5-number summary:

| min = 0.01 |

| 25% = 0.11 |

| 50% = 0.25 |

| 75% = 0.50 |

| max = 14.66 |

Name: Metrics.Sales, dtype: float64

- Για τη συγκεκριμένη μεταβλητή, και οι δύο τρόποι κάνουν καλή δουλειά στο να περιγράψουν την μεταβλητή, με τη σύνοψη των 5 αριθμών να είναι ίσως πιο κατάλληλη διότι υποδηλώνει το γεγονός ότι τα δεδομένα είναι σημαντικά δεξιά-στρεβλωμένα, με μια μέση τιμή που είναι λιγότερο ενδεικτική της κεντρικής τάσης λόγω της παρουσίας ακραίων σημείων με υψηλές τιμές (μέγιστο 14.66 έναντι 75% ποσοστού 0.50).

- -

- **Metrics.Used Price:**

Metrics.Used Price Average: 17.335021097046408

Metrics.Used Price Standard Deviation: 5.182464473041564

- Metrics.Used Price 5-number summary:

| min = 4.95 |

| 25% = 14.95 |

| 50% = 16.95 |

| 75% = 17.95 |

| max = 49.95 |

Name: Metrics.Used Price, dtype: float64

- Για τη συγκεκριμένη μεταβλητή, η σύνοψη των 5 αριθμών είναι η καταλληλότερη για να υπογραμμίσει το γεγονός ότι ενώ οι περισσότερες τιμές συγκεντρώνονται κάτω από $18, υπάρχει μια μεγάλη ποικιλία τιμών έως και $49.95, που μπορεί να μην είναι προφανές μόνο με τη μέση τιμή και την τυπική απόκλιση.

- -

- **<u>Release.Year:</u>**

Release.Year Average: 2006.746835443038

Release.Year Standard Deviation: 1.0595971153273522

- Release.Year 5-number summary:

| min = 2004 |

| 25% = 2006 |

| 50% = 2007 |

| 75% = 2008 |

| max = 2008 |

Name: Release.Year, dtype: int64

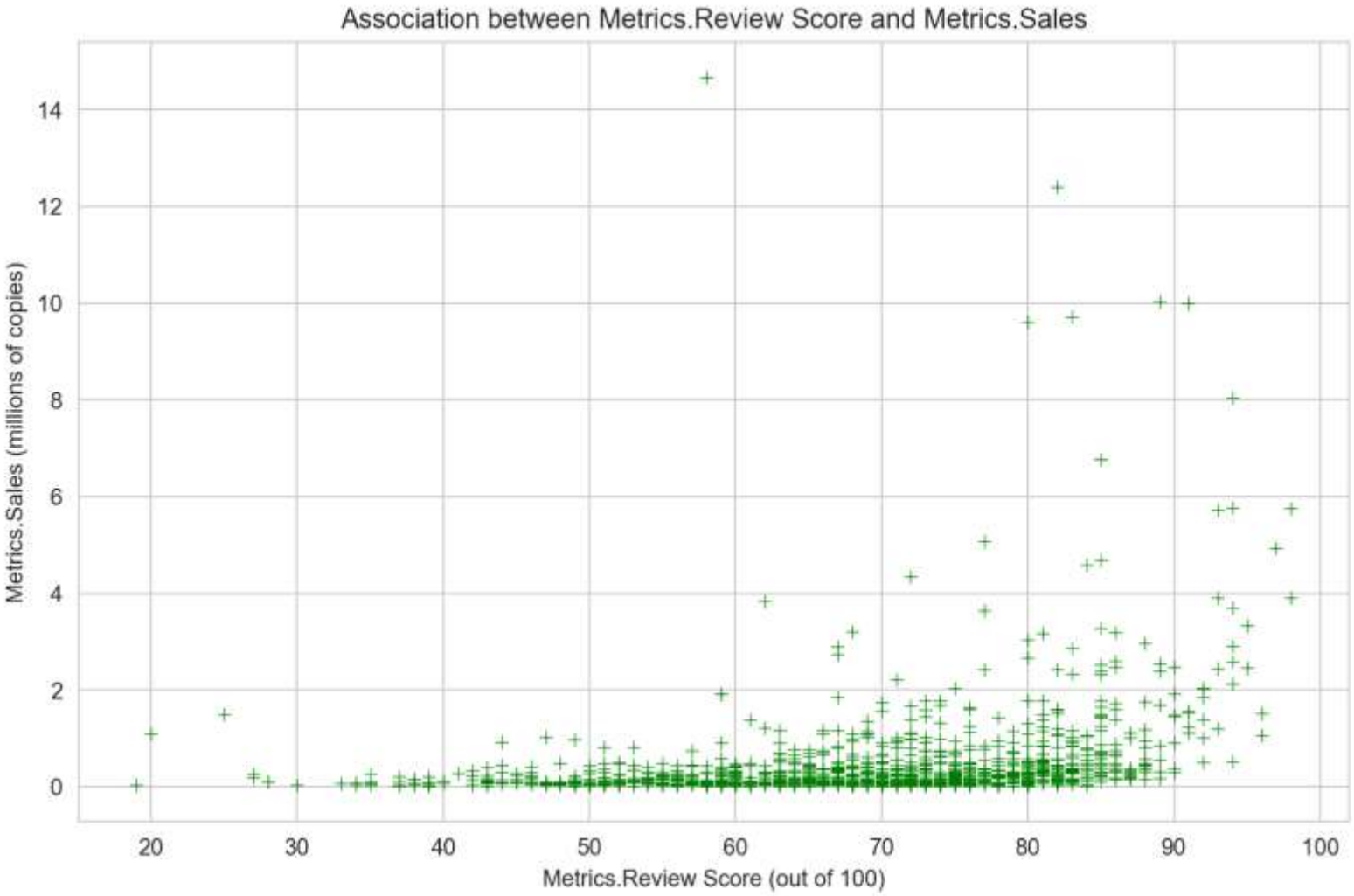
- Για τη συγκεκριμένη μεταβλητή, τα έτη είναι εντός ενός στενού εύρους και πιθανώς έχουν μια ομοιόμορφη ή κανονική κατανομή, έτσι η μέση τιμή και η τυπική απόκλιση είναι επαρκείς για να περιγράψουν την κεντρική τάση και τη διασπορά.

- -

ΝΙΚΟΛΑΟΣ ΜΗΤΣΑΚΗΣ [ΠΛΗΡ]
Α.Μ. : 3210122

e) Διερευνώ την σχέση μεταξύ των μεταβλητών: Review.Score και Metrics.Sales

Scatterplot:



Συμπέρασμα από το παραπάνω γράφημα:

Στο παραπάνω scatterplot παρατηρούμε ότι στο εύρος των σκορ από 70 έως 90 το πλήθος των σημείων είναι μεγάλο με το γράφημα να γίνεται εμφανέστατα πιο πυκνό.

Αυτό υποδηλώνει τα εξής:

- Τα περισσότερα βιντεοπαιχνίδια με υψηλές κριτικές (εύρος 70% με 100%) τείνουν να πουλάνε έως και 2 εκατομμύρια αντίγραφα, με λίγες εξαιρέσεις να φτάνουν ακόμη υψηλότερα νούμερα.

- Παρατηρούμε ότι το δείγμα μας δεν περιέχει πολλά βιντεοπαιχνίδια με χαμηλές κριτικές (κάτω από 30%) ή ότι γενικά δεν είναι συχνό φαινόμενο να υπάρχουν

βιντεοπαιχνίδια με χαμηλές κριτικές. Η προσωπική μου εμπειρία όμως, με κάνει να συγκλίνω προς το πρώτο ενδεχόμενο ως την κύρια αιτία.

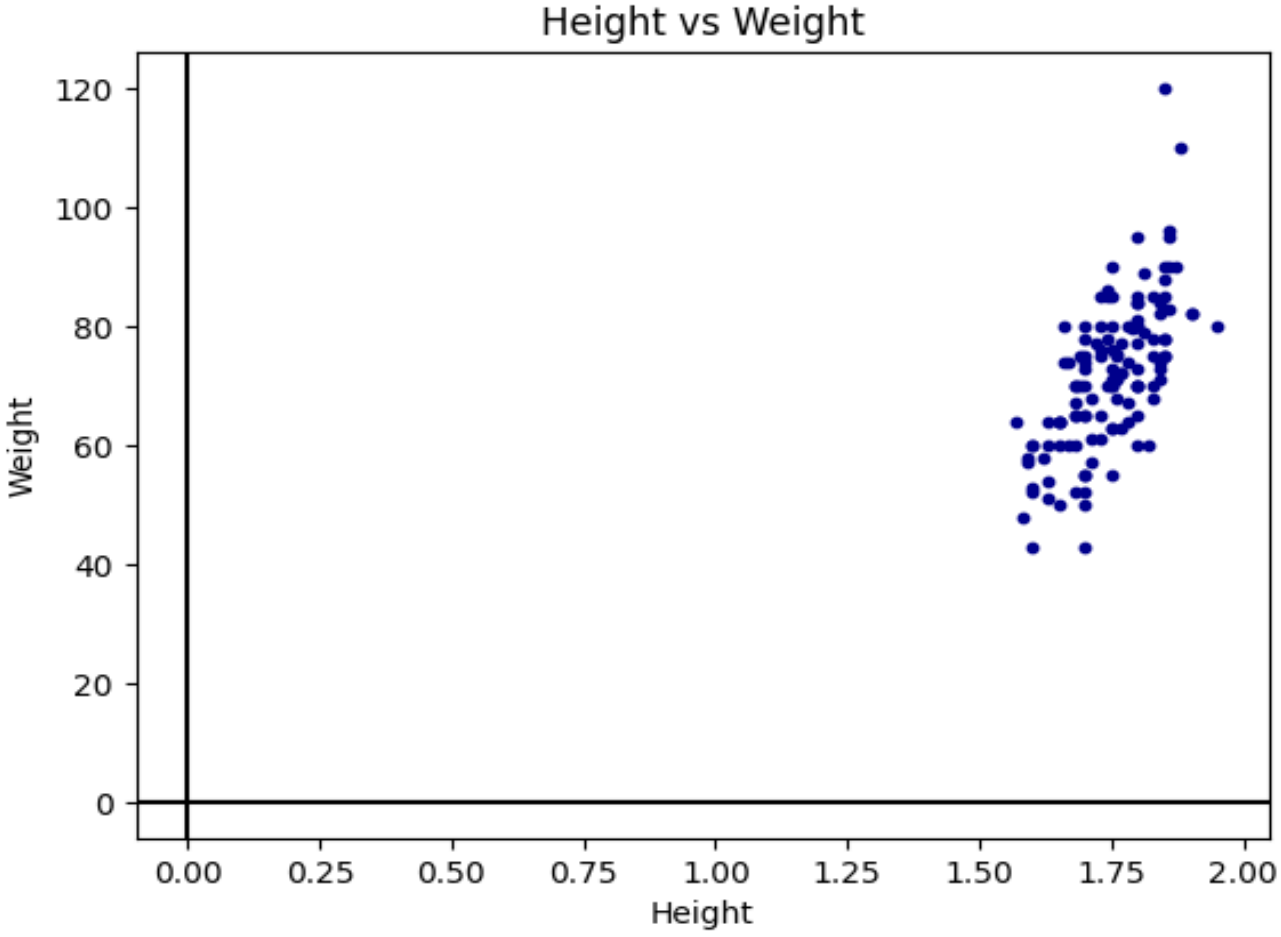
- Παρατηρούμε ότι ασχέτως με την κριτική του, ένα βιντεοπαιχνίδι κατά κύριο λόγο δεν θα πουλήσει πάνω από 2 εκατομμύρια αντίγραφα. Προφανώς και υπάρχουν, λίγες βέβαια, εξαιρέσεις παιχνιδιών που έχουν πουλήσει έως και 10 εκατομμύρια αντίγραφα με κριτική 80%.

- Δεν μπορούμε να διακρίνουμε κάποια ξεκάθαρη συσχέτιση μεταξύ των πωλήσεων και των κριτικών των βιντεοπαιχνιδιών. Δηλαδή, βιντεοπαιχνίδια με υψηλές κριτικές δεν υπόσχονται απαραίτητα και υψηλό αριθμό πωλήσεων.

Άσκηση 3).

a).

Εξετάζω την σχέση μεταξύ των μεταβλητών height και weight από το αρχείο ‘survey_data_2023.csv’ :



ΝΙΚΟΛΑΟΣ ΜΗΤΣΑΚΗΣ [ΠΛΗΡ]
Α.Μ. : 3210122

- Οι μεταβλητές 'Height' και 'Weight' είναι θετικά συσχετισμένες σε έντονο βαθμό, καθώς ο συντελεστής συσχέτισης είναι θετικός και ίσος με 0.67189, ένας αριθμός κοντά στο 1.

b).