ΑΡΙΣΤΟΤΕΛΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΟΝΙΚΗΣ
ΤΜΗΜΑ ΦΥΣΙΚΗΣ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
ΥΠΟΛΟΓΙΣΤΙΚΗΣ ΦΥΣΙΚΗΣ

ΤΜΗΜΑ ΦΥΣΙΚΗΣ Α.Π.Θ.
Πρόγραμμα Μεταπτυχιακών Σπουδών
Υπολογιστικής Φυσικής

Οπτικοποίηση των δεδομένων

Διακονίδης Θόδωρος
Ε.ΔΙ.Π
Εργαστήριο Θεωρ. Φυσικής
Τμήμα Φυσικής ΑΠΘ

Η σημασία της απεικόνισης

Η απεικόνιση είναι πολύ σημαντική για να αποκτηθεί μια πρώτη ιδέα του συνόλου δεδομένων.

Μπορούμε να εντοπίσουμε πολλές ανωμαλίες που θα πρέπει να φροντίσουμε, πριν συνεχίσουμε π.χ.

- Λάθη ή μη χρήσιμοι τύποι δεδομένων,
- τιμές που λείπουν

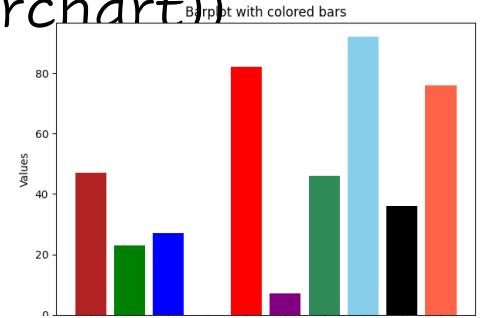
Ταυτόχρονα οι απεικονίσεις θα μας δώσουν πολύτιμες πληροφορίες σχετικά με τις μεταβλητές του συνόλου δεδομένων και τις συσχετίσεις μεταξύ τους.

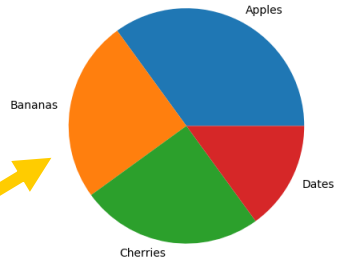
Τύποι μεταβλητών που θα συναντήσουμε

- **Ποσοτικές μεταβλητές**
  - **Συνεχείς:** Το σύνολο των δυνατών τιμών είναι ένα συνεχές υποσύνολο των πραγματικών αριθμών π.χ. ύψος, μέτρηση κάποιας φυσ. Ιδιότητας κλπ.
  - **Διακριτές:** Αποτελείται από συγκεκριμένες τιμές (συνήθως ακεραίους) π.χ. αριθμός γεννήσεων κλπ.
- **Ποιοτικές μεταβλητές**
  - Αναφέρονται σε κάποιο ποιοτικά διαχωρίσιμο χαρακτηριστικό π.χ. Ίαση-θάνατος, background-signal κλπ.

Απεικόνιση ποιοτικών (qualititative) μεταβλητών (Ραβδόγραμμα (Barchart))

- Μέτρηση των παρατηρήσεων που εμπίπτουν σε κάθε κατηγορία (συχνότητα), εναλλακτικά, το ποσοστό των παρατηρήσεων που αναλογεί σε κάθε κατηγορία (σχετική συχνότητα).





Εναλλακτικά μπορεί να απεικονιστούν και με διάγραμμα πίτας (pie-chart)

Βιβλιοθήκες για την απεικόνιση

- Θα χρησιμοποιήσουμε τις δυο βασικές βιβλιοθήκες απεικονίσεων, τόσο γι' αυτές όσο και για τις ποσοτικές μεταβλητές. Τις matplotlib και seaborn.

Εναλλακτικά μπορεί να κάποια/ος να χρησιμοποιήσει και τις ακόλουθες: **Bokeh** and **Altair**

Άσκηση 1

Με βάση αυτά που δείξαμε μέχρι τώρα. Σε ένα νέο jupyter notebook:

1. Εισάγετε το αρχείο sonar.all-data.csv χρησιμοποιώντας την βιβλιοθήκη pandas (Πληροφορίες για το dataset μπορείτε να βρείτε από εδώ: Sonar dataset info)
2. Ελέγξτε τα δεδομένα που αποτελείται και βρείτε τις ποιοτικές μεταβλητές.
3. Απεικονίστε σε ραβδόγραμμα και διάγραμμα πίτας τα αποτελέσματα της χρησιμοποιώντας τις δυο βιβλιοθήκες που αναφέραμε.
4. Τι παρατηρείτε;

Απεικόνιση ποσοτικών (quantitative) μεταβλητών

Θα δούμε διάφορους τύπους απεικόνισης και θα εξηγήσουμε την χρησιμότητα της κάθε μίας. Οι τύποι των διαγραμμάτων:

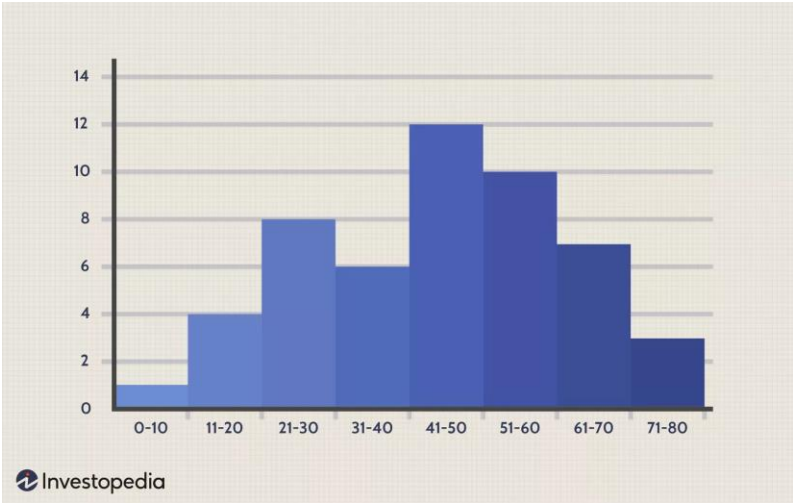
- Ιστογράμματα (Histograms)
- Θηκογράμματα (box plots)
- Violin plots
- Γραφήματα διασποράς (Scatter Plots)

Ιστογράμματα (Histograms)

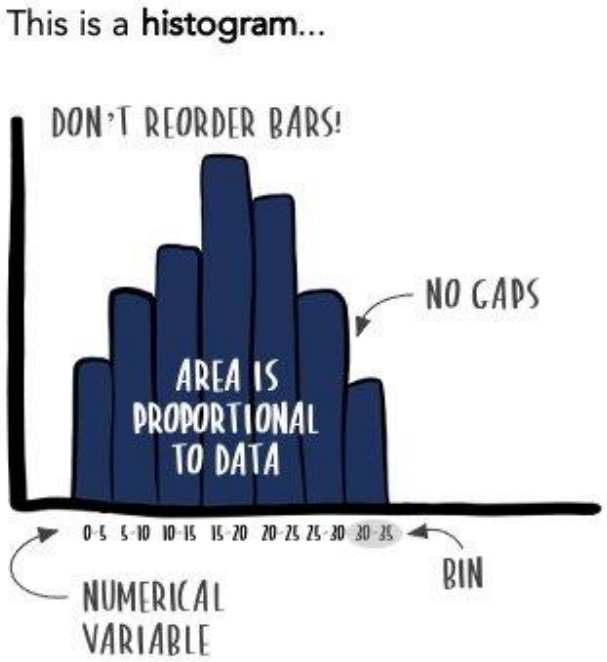
Γραφική απεικόνιση στατιστικών συχνοτήτων περιοχών τιμών ενός μεγέθους.

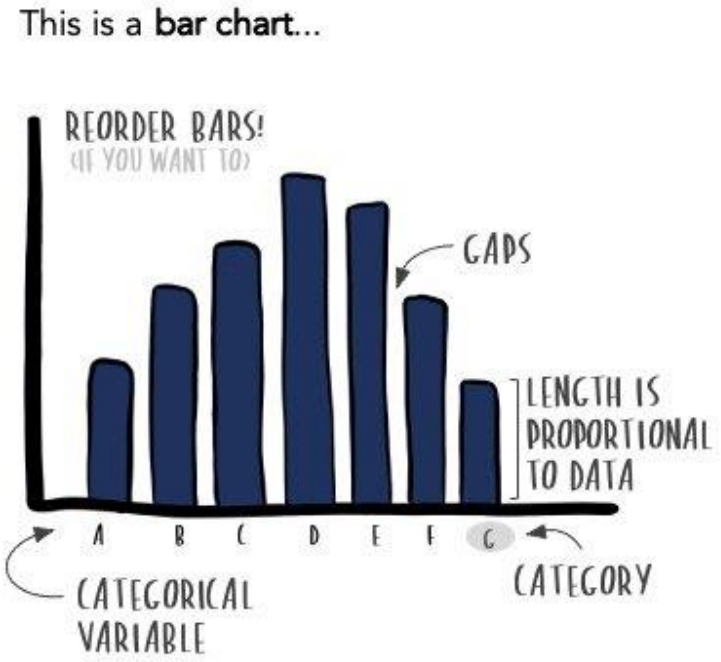
Στην πράξη:

- Ομαδοποίηση (σε bins) των συνεχών τιμών και διάταξη στον οριζόντιο άξονα κατ΄ αύξουσα σειρά.
- Κατασκευή ορθογωνίων, το ύψος των οποίων αντιστοιχεί στη συχνότητα (πλήθος) κάθε ομάδας.



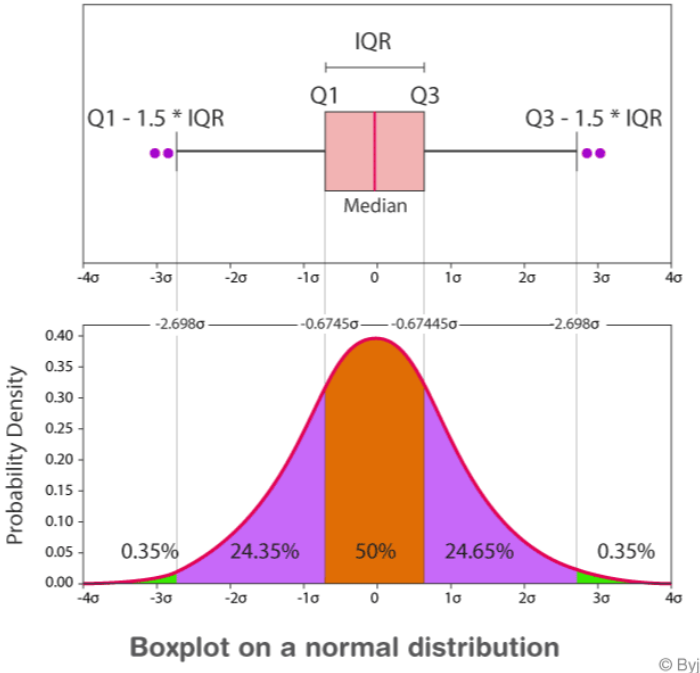
Διαφορές μεταξύ ιστογραμμάτων και ραβδογραμμάτων





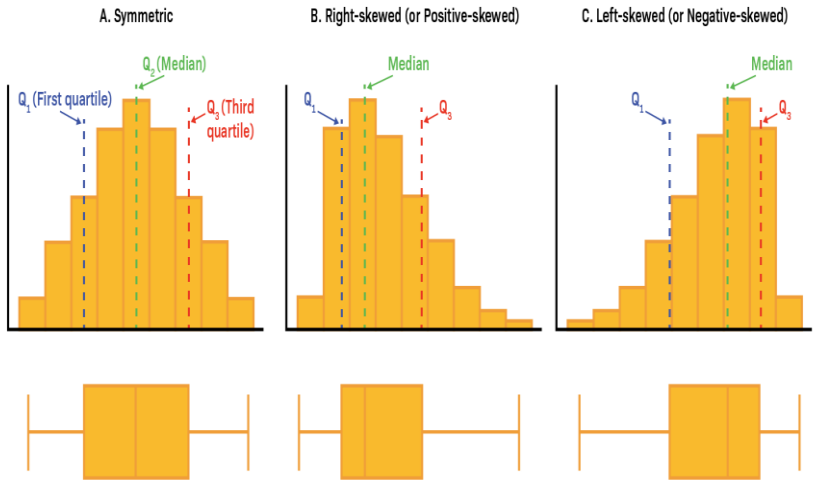
Θηκογράμματα (Box plots)

- Ονομάζεται και διάγραμμα 5 σημείων. Q1,Q3 τα 2 από τα 4 quartiles: 25%, 75% από το σύνολο το σημείων αντίστοιχα, είναι πριν από αυτά.
- Median (Ενδιάμεση τιμή) (Ταυτίζεται με το όριο του Q2).
- IQR: ενδοτεταρτομοριακό εύρος
- Σημεία πέρα του Q3+1,5*IQR και Q1-1,5IQR θεωρούνται ακραία (outliers)



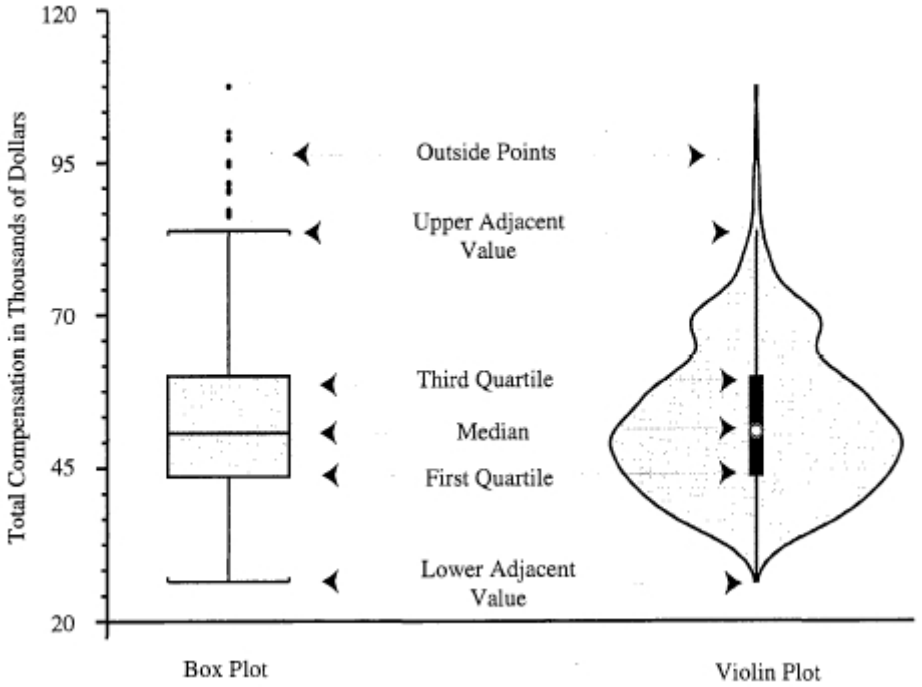
Θηκογράμματα (Box plots)

- Ιδιαίτερα χρήσιμα για την ανίχνευση ακραίων τιμών και την διαπίστωση συμμετρίας-ασυμμετρίας της κατανομής.
- Εξαιρετικά χρήσιμα για την σύγκριση κατανομών 2 ή περισσότερων δειγμάτων.
- Δε θεωρείται αξιόπιστο μέτρο διασποράς, επειδή βασίζεται μόνο στη μικρότερη και στη μεγαλύτερη παρατήρηση



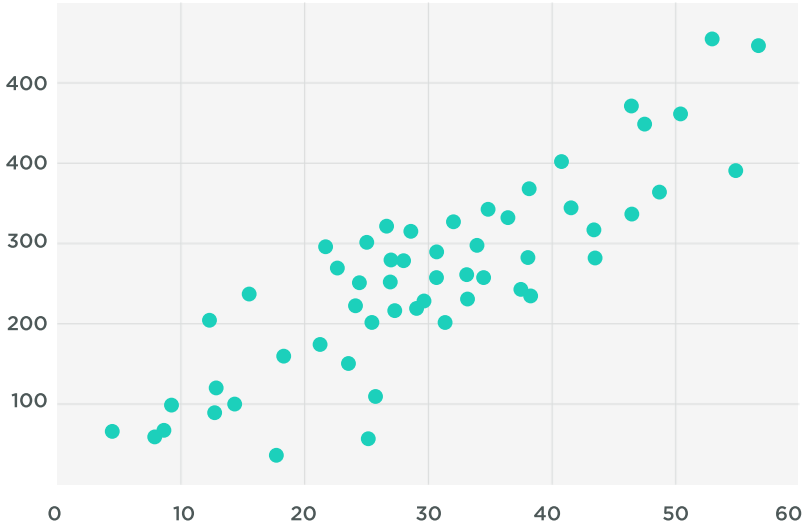
Violin plots

- Αποτελούν συμπλήρωμα των boxplots
- Καταγράφουν την διασπορά των δεδομένων, μαζί με τη σημαντική οπτική περιγραφή του θηκογράμματος
- Η προσθήκη του γραφήματος πυκνότητας (density graph) προσδίδει την έξτρα πληροφορία



Διαγράμματα Διασποράς (Scatter plots)

- Χρησιμοποιούνται για ανεύρεση πιθανής συσχέτισης μεταξύ 2 μεταβλητών.
- Τα δεδομένα εμφανίζονται ως σειρά σημείων, καθένα από τα οποία έχει την τιμή μιας μεταβλητής που καθορίζει τη θέση στον οριζόντιο άξονα και την τιμή της άλλης μεταβλητής που καθορίζει τη θέση στον κατακόρυφο άξονα.



Ελλιπή Δεδομένα (Missing data)

- Σε ένα dataset μπορούμε να έχουμε κενά ή μη οριζόμενους αριθμούς (NaN) ή ακόμη και μηδενικά που μεταφράζονται σαν έλλειψη δεδομένων.
- Δημιουργούν πρόβλημα στην περαιτέρω επεξεργασία
- Συνήθως συμπληρώνονται σπανιότερα (μη γενικώς προτεινόμενο) απαλείφονται.

	First Score	Second Score	Third Score	Fourth Score
0	100.0	30.0	52.0	60
1	NaN	NaN	NaN	67
2	NaN	45.0	80.0	68
3	95.0	56.0	98.0	65

Λύσεις που προτείνονται (Data Imputation)

- Αντιμετωπίζονται ανάλογα με την περίσταση.

Κάποιοι από τους τρόπους αντιμετώπισης:

1. Αντικατάστασή τους με μέση τιμή (mean) ή ενδιάμεση (median) της αντίστοιχης στήλης (μεταβλητής)
2. Με τη χρήση διαφόρων αλγορίθμων παραγωγής δεδομένων π.χ. Knn imputation (Βλ. και αλγόριθμο μηχανικής μάθησης)
3. Χρήση μεθόδου παλινδρόμησης: Περιλαμβάνει την πρόβλεψη των τιμών που λείπουν χρησιμοποιώντας ένα μοντέλο παλινδρόμησης, όπως γραμμική παλινδρόμηση ή δέντρα αποφάσεων κλπ. (Βλ. επιβλεπόμενη μηχανική μάθηση)

ΑΡΙΣΤΟΤΕΛΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΟΝΙΚΗΣ
ΤΜΗΜΑ ΦΥΣΙΚΗΣ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
ΥΠΟΛΟΓΙΣΤΙΚΗΣ ΦΥΣΙΚΗΣ

ΤΜΗΜΑ ΦΥΣΙΚΗΣ Α.Π.Θ.
Πρόγραμμα Μεταπτυχιακών Σπουδών
Υπολογιστικής Φυσικής

Διαχείριση δεδομένων

Διακονίδης Θόδωρος
Ε.ΔΙ.Π
Εργαστήριο Θεωρ. Φυσικής
Τμήμα Φυσικής ΑΠΘ

Διάφοροι τύποι μετασχηματισμού δεδομένων

Αυτοί που θα χρησιμοποιήσουμε για να επεξεργαστούμε τα δικά μας δεδομένα και οι πιο διαδεδομένες είναι οι:

1. **Κανονικοποίηση μεγίστου ελαχίστου (Min-max normalization)**

Μετατρέπει κάθε τιμή των δεδομένων σε μια περιοχή μεταξύ 0 και 1.

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}}$$

2. **Standardization Z-score**

Μετατρέπει κάθε τιμή των δεδομένων αφαιρώντας την από τον **μέσο όρο** του δείγματος και διαιρώντας στη συνέχεια με την **τυπική απόκλιση** του δείγματος. Δεν έχει προκαθορισμένο μέγιστο ή ελάχιστο.

$$x' = \frac{x - \mu}{\sigma}$$

Ανάλυση Βασικών Συνιστωσών (Principal Component Analysis, PCA)

Τι ακριβώς είναι το PCA;

Είναι μια μέθοδος μείωσης διαστάσεων των μεγάλων συνόλων δεδομένων, μετατρέποντας τα σε μικρότερα, τα οποία εξακολουθούν να περιέχουν τις περισσότερες πληροφορίες του μεγάλου συνόλου.

Που μπορεί να χρησιμοποιηθεί;

Στην ανάλυση δεδομένων πολλών διαστάσεων, αραιού καταμερισμού (big sparsity) ή δεδομένων που υπάρχει υποψία ότι συσχετίζονται μεταξύ τους. Η διαδικασία αυτή οδηγεί σε συμπύκνωση της πληροφορίας και επομένως στην καλύτερη οπτικοποίηση και ταχύτερη ανάλυση αυτών, χωρίς την ανάγκη ανάλυσης επιπλέον μεταβλητών.

Πως ακριβώς επιτυγχάνεται ο καταμερισμός της πληροφορίας

- Αυτό ουσιαστικά που επιτυγχάνεται είναι η επιλογή καταλλήλου άξονα, η προβολή στον οποίο των δεδομένων να οδηγεί στην μεγαλύτερη διασπορά (variance), που μεταφράζεται σε μεγαλύτερο ποσοστό της πληροφορίας των δεδομένων.

- Ακολουθεί ο κάθετος σε αυτόν (δύο διαστάσεις), έτσι δημιουργούνται 2 PC1 και PC2 με την μεγαλύτερη πληροφορία να υπάρχει στον πρώτο.