

Μέρος III – Οπτικοποίηση Δεδομένων

6. Δέκα βασικές αρχές

Η οπτικοποίηση δεδομένων δεν είναι μια ουδέτερη τεχνική διαδικασία, αλλά μια μορφή επικοινωνίας. Κάθε γράφημα ενσωματώνει επιλογές, όπως τι θα παρουσιαστεί, τι θα τονιστεί, τι θα παραλειφθεί και πού να εστιάσει ο αναγνώστης. Οι επιλογές αυτές επηρεάζουν άμεσα την κατανόηση των δεδομένων και τα συμπεράσματα που εξάγονται.

Παρά την ευρεία διαθεσιμότητα εργαλείων οπτικοποίησης, η παραγωγή αποτελεσματικών γραφημάτων δεν είναι αυτονόητη. Συχνά, γραφήματα που είναι τεχνικά ορθά αποτυγχάνουν να μεταδώσουν το μήνυμα. Για παράδειγμα, μπορεί να είναι υπερφορτωμένα, να παραπλανούν, ή απλώς να μην είναι προσαρμοσμένα στο κοινό στο οποίο απευθύνονται. Επομένως, η οπτικοποίηση απαιτεί τόσο τεχνικές δεξιότητες όσο και κατανόηση βασικών αρχών σχεδιασμού.

Στο κεφάλαιο αυτό παρουσιάζονται δέκα θεμελιώδεις αρχές οπτικοποίησης, οι οποίες συνοψίζουν βέλτιστες πρακτικές από τη βιβλιογραφία και την εμπειρία της επιστημονικής κοινότητας. Οι αρχές αυτές δεν αποτελούν αυστηρούς κανόνες, αλλά κατευθυντήριες γραμμές που βοηθούν στη λήψη συνειδητών σχεδιαστικών αποφάσεων. Καλύπτουν ζητήματα όπως η κατανόηση του κοινού, η σαφής διατύπωση του μηνύματος, η αποφυγή παραπλανητικών απεικονίσεων και η αποτελεσματική χρήση χρώματος και μορφής.

Στόχος του κεφαλαίου είναι να αναδείξει ότι η επιτυχία μιας οπτικοποίησης δεν εξαρτάται πρωτίστως από το εργαλείο που χρησιμοποιείται, αλλά από τον τρόπο με τον οποίο σχεδιάζεται. Οι αρχές που παρουσιάζονται εδώ αποτελούν το εννοιολογικό υπόβαθρο πάνω στο οποίο θα στηριχθεί το επόμενο κεφάλαιο, όπου εξετάζονται συγκεκριμένες βιβλιοθήκες και τεχνικές υλοποίησης.

6.1. Αρχή I: Επίγνωση του αποδέκτη

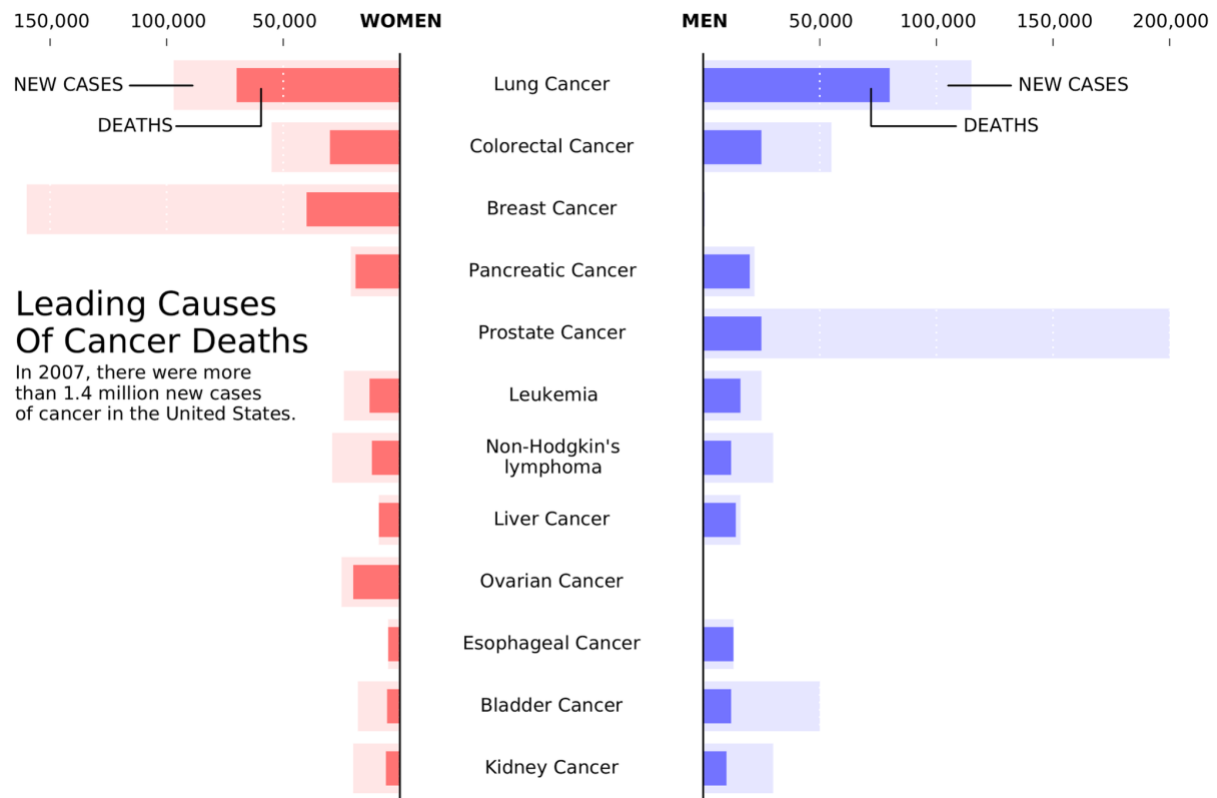
Η πρώτη και θεμελιώδης αρχή της οπτικοποίησης δεδομένων αφορά τον παραλήπτη του οπτικοποιημένου μηνύματος. Κάθε γράφημα είναι μια πράξη *επικοινωνίας* και αποκτά νόημα μόνο σε σχέση με το κοινό στο οποίο απευθύνεται.

Η ίδια πληροφορία μπορεί να παρουσιαστεί με εντελώς διαφορετικό τρόπο ανάλογα με το επίπεδο γνώσης και τον σκοπό. Ένα γράφημα που προορίζεται για επιστήμονες μπορεί να περιλαμβάνει αναλυτικές πληροφορίες, τεχνικούς όρους, περίπλοκες καμπύλες και λεπτομερείς άξονες. Το ίδιο γράφημα, αν απευθύνεται στο ευρύ κοινό, θα πρέπει να είναι απλό, με λίγες μεταβλητές, καθαρές ετικέτες και να είναι ρητό το βασικό συμπέρασμα.

Η αποτυχία προσαρμογής στο κοινό οδηγεί συχνά σε δύο αντίθετα σφάλματα. Στην πρώτη περίπτωση, η οπτικοποίηση είναι υπερβολικά σύνθετη (π.χ., φορτωμένη με τεχνικές λεπτομέρειες που αποπροσανατολίζουν από το κεντρικό μήνυμα). Στη δεύτερη, είναι υπεραπλουστευμένη (δηλαδή, αφαιρεί κρίσιμες πληροφορίες, πιθανώς παραπλανώντας).

Το γράφημα της Εικόνας 6.1 αποτελεί ανακατασκευή μιας οπτικοποίησης που είχε δημοσιευθεί στους *The New York Times* το 2007. Η εκδοχή αυτή, που παρουσιάστηκε σε ένα

επιστημονικό άρθρο,⁵ δημιουργήθηκε με matplotlib και βασίζεται σε προσεγγιστικά δεδομένα. Περιλαμβάνει τέσσερις σειρές, για θανάτους και περιστατικά καρκίνου για άνδρες και γυναίκες. Θα μπορούσαν να παρουσιαστούν με έναν «κλασικό» διπλό ραβδόγραμμα (θάνατοι/νέα περιστατικά). Ωστόσο, οι συγγραφείς επιλέγουν διαφορετική διάταξη.



Εικόνα 6.113 Ανακατασκευή οπτικοποίησης των Rougier et al. (2014) με άδεια CC0 που είχε δημοσιευθεί στους New York Times το 2007 για τις αιτίες θανάτων από καρκίνο σε γυναίκες (κόκκινο χρώμα, αριστερά) και τους άνδρες (μπλε χρώμα, δεξιά).

Η επιλογή αυτή δεν είναι αισθητική. Η διάταξη εκμεταλλεύεται το γεγονός ότι τα νέα περιστατικά είναι πάντοτε περισσότερα από τους αντίστοιχους θανάτους, επιτρέποντας τη συνδυαστική απεικόνιση των δύο μεγεθών σε μία ενιαία οπτική δομή. Παράλληλα, αξιοποιεί τη φυσική κατεύθυνση ανάγνωσης στα αγγλικά (αριστερά προς τα δεξιά), διευκολύνοντας τη σύγκριση ανδρών και γυναικών. Οι κεντρικές ετικέτες καθιστούν άμεσα προσβάσιμο το βασικό θέμα που αφορά θανάτους από καρκίνο. Το γράφημα, επίσης, είναι «αυτοτελές» μεταδίδοντας το κύριο νόημα χωρίς να απαιτείται εκτενής συνοδευτική ανάλυση.

Και όμως, η ίδια επιλογή που το καθιστά αποτελεσματικό για το ευρύ κοινό το καθιστά προβληματικό για επιστημονική χρήση. Η συγκεκριμένη διάταξη δυσκολεύει την ακριβή εκτίμηση αριθμητικών τιμών, ιδίως για κατηγορίες που βρίσκονται χαμηλά (όπως ο καρκίνος του νεφρού), ενώ οι σημάνσεις των αξόνων βρίσκονται ψηλά και δεν υπάρχουν βοηθητικές εσωτερικές γραμμές. Το γράφημα δεν σχεδιάστηκε για ακριβή ποσοτική ανάγνωση, αλλά για την άμεση κατανόηση της σχετικής βαρύτητας των κατηγοριών που πραγματεύεται.

⁵ Rougier, N. P., Droettboom, M., & Bourne, P. E. (2014). Ten simple rules for better figures. *PLoS computational biology*, 10(9), e1003833.

Για μια εφημερίδα γενικού ενδιαφέροντος, η έμφαση στο μήνυμα υπερισχύει της αριθμητικής ακρίβειας. Για ένα επιστημονικό άρθρο, η έλλειψη ακριβών τιμών θα ήταν μη αποδεκτή (εκτός αν αυτές παρέχονταν αλλού). Η διάκριση αυτή είναι κρίσιμη για την επιστήμη δεδομένων. Η οπτικοποίηση που απευθύνεται σε ερευνητές οφείλει να επιτρέπει αναλυτικό έλεγχο και ποσοτική εκτίμηση. Η οπτικοποίηση που απευθύνεται σε μη ειδικούς οφείλει να αναδεικνύει καθαρά το βασικό συμπέρασμα. Και οι δύο μπορεί να είναι ορθές, αρκεί να είναι **σχεδιασμένες με επίγνωση του αποδέκτη**.

Ο πρώτος κανόνας, λοιπόν, δεν προτείνει απλώς προσαρμογή του ύφους. Υποδεικνύει ότι κάθε σχεδιαστική επιλογή, από τη διάταξη, την κλίμακα, την τοποθέτηση ετικετών, ως και την κατεύθυνση ανάγνωσης, πρέπει να ερμηνεύεται ως μία επικοινωνιακή απόφαση. Το ερώτημα δεν είναι αν «είναι το γράφημα σωστό» αλλά αν «είναι σωστό ως προς ποιον».

6.2. Αρχή II: Προσδιορισμός του μηνύματος

Αν ο πρώτος κανόνας αφορά το **σε ποιον μιλάω**, ο δεύτερος αφορά το **τι θέλω να πω**. Συχνά, όταν εργαζόμαστε με δεδομένα, μπαίνουμε στον πειρασμό να δείξουμε τα πάντα, όλες τις μεταβλητές, όλες τις σχέσεις, όλες τις πιθανές ερμηνείες. Το αποτέλεσμα είναι ένα γράφημα που περιέχει πληροφορία αλλά όχι κατεύθυνση. Ο αναγνώστης βλέπει τα δεδομένα, αλλά δεν γνωρίζει τι πρέπει να παρατηρήσει.

Ο δεύτερος κανόνας μάς καλεί να αναρωτηθούμε πριν από κάθε σχεδιαστική απόφαση ποιο είναι το βασικό συμπέρασμα που θέλουμε να μείνει στον αναγνώστη. Μπορεί να είναι μία τάση στον χρόνο, μία διαφορά μεταξύ ομάδων, ή μία ανωμαλία. Το μήνυμα, όμως, διαφέρει από την περιγραφή. Το «τα περιστατικά αυξάνονται» είναι μία περιγραφή. Το «η αύξηση είναι εντονότερη μετά το 2015» είναι ένα μήνυμα. Η διαφορά είναι ότι το δεύτερο κατευθύνει την προσοχή.

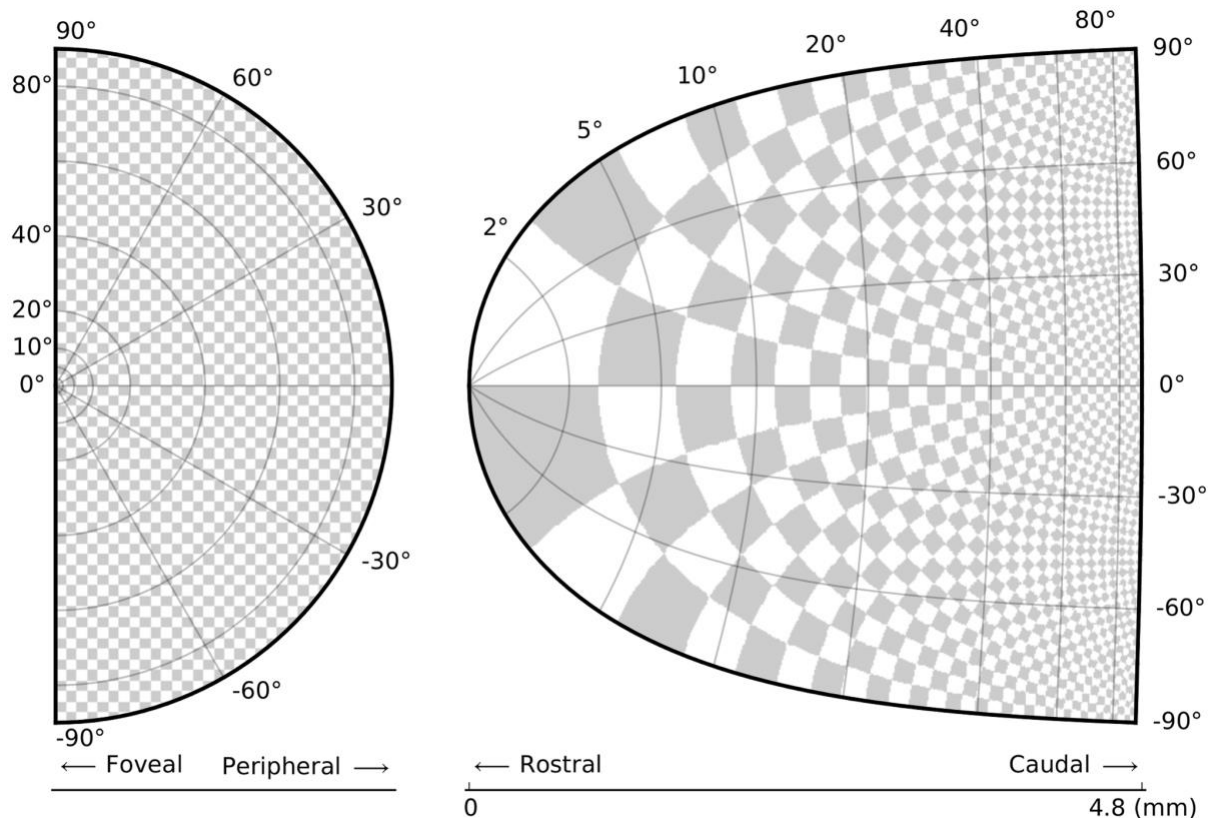
Η σαφήνεια του μηνύματος επηρεάζει τη μορφή του γραφήματος. Αν το ζητούμενο είναι η σύγκριση δύο ομάδων, ο σχεδιασμός πρέπει να διευκολύνει τη σύγκριση, όχι να την αποκρύπτει μέσα σε διάφορες καμπύλες. Αν το ζητούμενο είναι μια μεταβολή στον χρόνο, η χρονική διάσταση πρέπει να κυριαρχεί οπτικά. Αν το ζητούμενο είναι η κατανομή, τότε η μορφή του γραφήματος πρέπει να αναδεικνύει το σχήμα της. Η έλλειψη σαφούς μηνύματος, ακόμη και αν μοιάζει με ουδετερότητα, συχνά οδηγεί σε υπερφόρτωση πληροφορίας, γιατί το γράφημα μετατρέπεται σε μία άτυπη αποθήκη δεδομένων.

Η αναγνώριση του μηνύματος προηγείται του σχεδιασμού. Δεν επιλέγουμε πρώτα τύπο γραφήματος και μετά αναζητούμε τι «λέει». Αντίθετα, διατυπώνουμε πρώτα το συμπέρασμα σε μία πρόταση και στη συνέχεια σχεδιάζουμε το γράφημα έτσι ώστε να το καθιστά εμφανές. Ο δεύτερος κανόνας μάς υπενθυμίζει ότι η οπτικοποίηση δεν είναι ουδέτερη τεχνική διαδικασία. Είναι πράξη εστίασης. Και η εστίαση ξεκινά από την απόφαση για το τι αξίζει να φανεί καθαρά.

Μια πρακτική άσκηση για να φανεί η αποτελεσματικότητα του γραφήματος είναι να γράψουμε το μήνυμα σε μία γραμμή και να ελέγξουμε αν κάποιος που βλέπει μόνο το γράφημα μπορεί να το συμπεράνει.

Το παράδειγμα της Εικόνας 6.2 αφορά μια δομή του εγκεφαλικού στελέχους (superior colliculus) που εμπλέκεται στον έλεγχο των κινήσεων των οφθαλμών. Νευροφυσιολογικές

μελέτες υποστηρίζουν ότι ο πληθυσμός ενεργών νευρώνων στην περιοχή αυτή κωδικοποιεί τη θέση ενός οπτικού στόχου. Η χαρτογράφηση από τον αμφιβληστροειδή στην επιφάνεια του colliculus βασίζεται σε ένα ποσοτικό μοντέλο **λογαριθμικής** απεικόνισης, μια μαθηματική συνάρτηση που μετατρέπει τις συντεταγμένες του αμφιβληστροειδούς σε συντεταγμένες της επιφάνειας του colliculus.



Εικόνα 6.214 Χαρτογράφηση από τον αμφιβληστροειδή στην επιφάνεια του *super colliculus*, η οποία βασίζεται σε ένα μοντέλο λογαριθμικής απεικόνισης. Από Rougier et al. (2014), Άδεια CC0.

Το μήνυμα της οπτικοποίησης, όμως, δεν είναι απλώς ότι *υπάρχει λογαριθμική χαρτογράφηση*. Το ουσιώδες μήνυμα είναι ότι αυτή η χαρτογράφηση οδηγεί σε έντονη μεγέθυνση της περιοχής της ωχράς κηλίδας (fovea). Για να καταστεί αυτό άμεσα αντιληπτό, οι συγγραφείς χρησιμοποιούν ένα τεχνητό μοτίβο σκακιέρας. Παρότι τέτοιο μοτίβο δεν χρησιμοποιείται σε πραγματικά πειράματα, λειτουργεί ως οπτικό εργαλείο που αναδεικνύει καθαρά τη γεωμετρική παραμόρφωση. Η σκακιέρα καθιστά ορατή τη δυσανάλογη μεγέθυνση της κεντρικής περιοχής, υπηρετώντας το βασικό μήνυμα.

Εδώ φαίνεται καθαρά η προτεραιότητα του νοήματος έναντι της κυριολεκτικής πιστότητας. Το γράφημα δεν επιχειρεί να αναπαραστήσει ένα ακριβές πειραματικό ερέθισμα. Επιλέγει μια τεχνητή, αλλά διδακτικά αποτελεσματική αναπαράσταση, προκειμένου να φωτίσει την κεντρική ιδέα. Η σχεδιαστική απόφαση καθοδηγείται από το μήνυμα.

Η αρχή αυτή έχει γενική ισχύ. Μόνο αφού διατυπώσουμε με σαφήνεια τι θέλουμε να κατανοήσει ο αναγνώστης αξίζει να επενδύσουμε χρόνο στον σχεδιασμό του γραφήματος. Όπως στη συγγραφή, δεν ξεκινάμε να διαμορφώνουμε προτάσεις πριν αποφασίσουμε τα βασικά σημεία του κειμένου, έτσι και στην οπτικοποίηση δεν επιλέγουμε μορφή πριν αποσαφηνίσουμε τον πυρήνα του επιχειρήματος. Όταν ένα γράφημα καταφέρνει να μεταδώσει ένα ισχυρό μήνυμα με την πρώτη ματιά, αυξάνονται οι πιθανότητες το άρθρο να

προσελκύσει την προσοχή της επιστημονικής κοινότητας. Η σαφήνεια δεν αφορά μόνο την αισθητική αλλά δίνει και στρατηγικό πλεονέκτημα στην επιστημονική επικοινωνία.

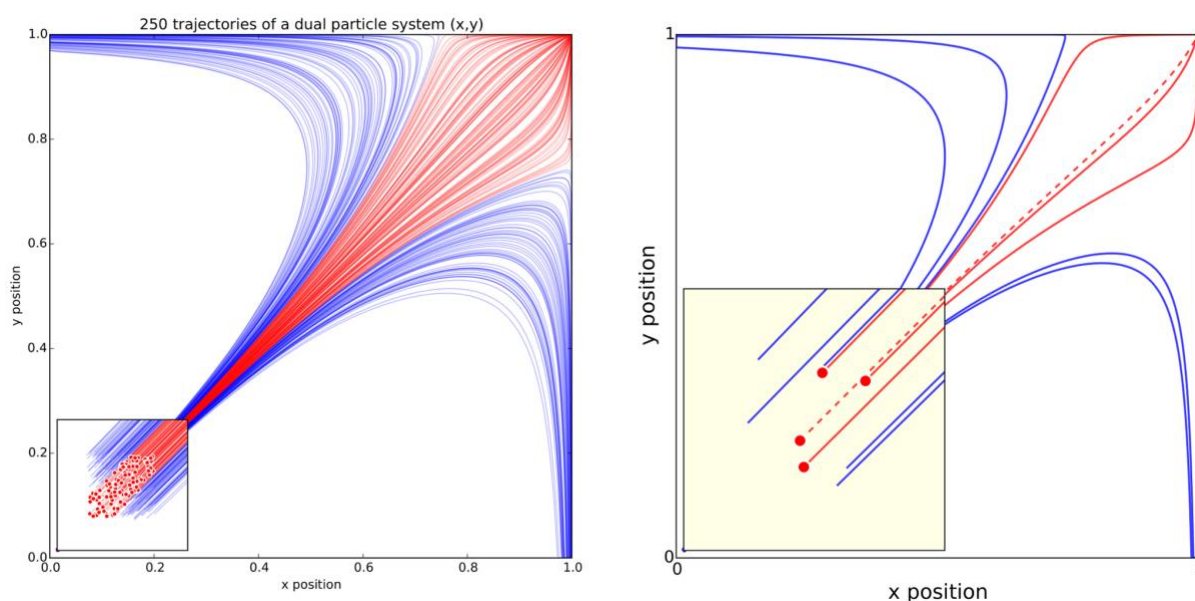
Ο δεύτερος κανόνας, επομένως, μας υπενθυμίζει ότι κάθε οπτικοποίηση είναι απάντηση σε ένα συγκεκριμένο ερώτημα. Αν δεν γνωρίζουμε ποιο είναι αυτό το ερώτημα, καμία τεχνική δεξιότητα δεν μπορεί να διασώσει το αποτέλεσμα.

6.3. Αρχή III: Προσαρμογή στο μέσο

Μια οπτικοποίηση δεν υπάρχει εκτός πλαισίου. Προβάλλεται, εκτυπώνεται, μεγεθύνεται, προβάλλεται από απόσταση ή εξετάζεται προσεκτικά σε κάποια οθόνη (π.χ., υπολογιστή). Το μέσο δεν είναι ουδέτερο. Αντίθετα, επηρεάζει ριζικά τον τρόπο με τον οποίο γίνεται αντιληπτό το γράφημα. Ο τρίτος κανόνας μάς υπενθυμίζει ότι κάθε μέσο συνεπάγεται διαφορετικούς περιορισμούς, αλλά και διαφορετικές δυνατότητες.

Το ίδιο γράφημα μπορεί να εμφανιστεί σε αφίσα, σε διαφάνεια συνεδρίου, σε έντυπο άρθρο ή σε ηλεκτρονική δημοσίευση. Σε μια προφορική παρουσίαση, ο χρόνος προβολής είναι περιορισμένος. Το κοινό πρέπει να κατανοήσει γρήγορα τι απεικονίζεται, ενώ ταυτόχρονα ακούει τον ομιλητή. Η πληροφορία πρέπει να είναι άμεσα προσβάσιμη, οπτικά ιεραρχημένη και καθαρή. Οι γραμμές χρειάζεται να είναι παχύτερες, τα σημεία και τα γράμματα μεγαλύτερα, η αντίθεση έντονη. Κάθε στοιχείο που απαιτεί παρατεταμένη παρατήρηση λειτουργεί εις βάρος της κατανόησης.

Σε ένα επιστημονικό άρθρο, από την άλλη πλευρά, ο αναγνώστης έχει τη δυνατότητα να αφιερώσει χρόνο. Μπορεί να εξετάσει λεπτομέρειες, να συγκρίνει περιοχές, να διαβάσει προσεκτικά το υπόμνημα και τη λεζάντα. Σε αυτό το πλαίσιο είναι λογικό να επιτρέπεται μεγαλύτερη πυκνότητα πληροφορίας. Επιπλέον επεξηγήσεις μπορούν να δοθούν στη λεζάντα, ενώ οι λεπτομέρειες αποκτούν νόημα.



Εικόνα 6.315 Προσομοίωση ενός συστήματος δύο σωματιδίων σε δύο εκδοχές. Η αριστερή εκδοχή έχει όλες τις πληροφορίες, ενώ στη δεξιά έχουν αφαιρεθεί λεπτομέρειες χάριν της προφορικής παρουσίασης. Από Rougier et al. (2014), Άδεια CC0.

Η Εικόνα 6.3 παρουσιάζει το αποτέλεσμα προσομοίωσης ενός συστήματος δύο σωματιδίων, ακριβώς το ίδιο αλλά σε δύο διαφορετικές εκδοχές. Η αριστερή εκδοχή είναι σχεδιασμένη

για το επιστημονικό άρθρο και διατηρούνται πολλές τροχιές. Οι γραμμές είναι ημιδιάφανες ώστε να αναδεικνύονται περιοχές υψηλής πυκνότητας και υπάρχει συνέπεια στη χρήση του κόκκινου χρώματος για τις αρχικές συνθήκες και τις τροχιές. Η πληροφορία είναι πλούσια και επιτρέπει διεξοδική εξέταση.

Η δεξιά εκδοχή είναι σχεδιασμένη για προφορική παρουσίαση ερευνητικών αποτελεσμάτων, μέσω κάποιου προβολέα σε κοινό που θα παρακολουθεί σε μία πολύ μεγάλη αίθουσα. Έτσι, έχουν αφαιρεθεί λεπτομέρειες, ο αριθμός των τροχιών μειώνεται, οι επικαλύψεις αποφεύγονται, τα σημεία αναφοράς περιορίζονται, οι άξονες και τα γράμματα μεγαλώνουν, οι γραμμές γίνονται παχύτερες. Επιπλέον, τροποποιούνται στοιχεία, όπως οι διακεκομμένες γραμμές, ώστε να είναι εύκολη η προφορική αναφορά τους. Το γράφημα δεν επιδιώκει πλήρη τεχνική πληρότητα, αλλά μία αποτελεσματική επικοινωνία σε περιορισμένο χρόνο.

Το δεξί γράφημα εκμεταλλεύεται τη σύγχρονη ανάγνωση σε ψηφιακές οθόνες. Οι αναγνώστες μπορούν να μεγεθύνουν, να μετακινήσουν και να εξερευνήσουν ένα σχήμα. Στα πλαίσια της προφορικής παρουσίασης, λοιπόν, μπορεί να γίνει προσομοίωση αυτής της ανάγνωσης.

Ο τρίτος κανόνας, γενικότερα, μάς καλεί να εγκαταλείψουμε την πρακτική της αυτούσιας μεταφοράς ενός σχήματος από το άρθρο στις διαφάνειες. Ιδανικά, κάθε μέσο απαιτεί ξεχωριστό σχεδιασμό. Το γράφημα δεν είναι ένα στατικό αντικείμενο, αλλά ένα προσαρμοζόμενο εργαλείο επικοινωνίας.

6.4. Αρχή IV: Η λεζάντα δεν είναι προαιρετική

Καμία οπτικοποίηση δεν είναι πλήρης από μόνη της. Όσο προσεκτικά κι αν έχει σχεδιαστεί, ένα γράφημα δεν μπορεί να εξηγήσει όλα όσα απαιτούνται για την πλήρη κατανόησή του. Γι' αυτό κάθε σχήμα πρέπει να συνοδεύεται από λεζάντα (caption), η οποία αποτελεί ένα αναπόσπαστο μέρος της επιστημονικής επικοινωνίας και όχι ένα τυπικό συμπλήρωμα.

Ένα γράφημα μπορεί να παρουσιάζει μια πειραματική διάταξη, ένα θεωρητικό μοντέλο ή νέα αποτελέσματα. Ωστόσο, δεν είναι εφικτό να ενσωματωθούν όλες οι επεξηγήσεις μέσα στο γράφημα. Δεν είναι καν επιθυμητό, γιατί η υπερφόρτωση με κείμενο καταστρέφει τη σαφήνεια. Αντίθετα, η λεζάντα αναλαμβάνει να εξηγήσει πώς «διαβάζεται» το γράφημα, να προσθέσει λεπτομέρειες που δεν μπορούν να αποδοθούν γραφικά και να αποσαφηνίσει όρους, παραδοχές ή παραμέτρους. Στην Εικόνα 6.3, για παράδειγμα, αναφέρεται η άδεια.

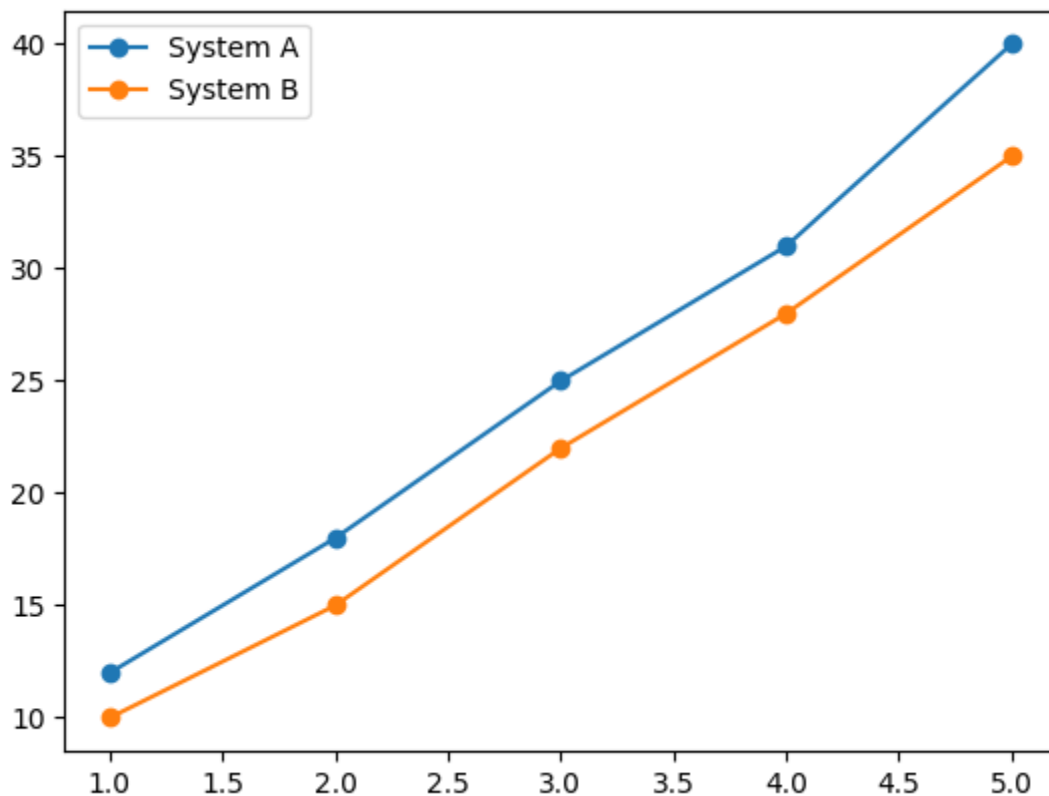
Μπορούμε να σκεφτούμε τη λεζάντα ως την εξήγηση που θα δίναμε προφορικά μπροστά από μια διαφάνεια ή μια αφίσα, με τη διαφορά ότι εδώ πρέπει να προβλέψουμε εκ των προτέρων τις πιθανές απορίες του αναγνώστη. Η λεζάντα οφείλει να απαντά σε ερωτήματα όπως: Τι ακριβώς απεικονίζεται; Ποιες είναι οι μονάδες; Ποια δεδομένα χρησιμοποιήθηκαν; Υπάρχουν μετασχηματισμοί ή κανονικοποιήσεις;

Ας δούμε ένα παράδειγμα με κώδικα, όπου συγκρίνονται δύο συστήματα:

```
import matplotlib.pyplot as plt
import numpy as np

# Δημιουργία ενδεικτικών δεδομένων
x = np.arange(1, 6)
system_1 = [12, 18, 25, 31, 40]
system_2 = [10, 15, 22, 28, 35]
```

```
plt.figure()
plt.plot(x, system_1, marker='o', label='System A')
plt.plot(x, system_2, marker='o', label='System B')
```



Εικόνα 6.416 Επίδοση δύο συστημάτων σε ένα γράφημα όπου λείπουν πολλές ουσιαστικές πληροφορίες.

Το γράφημα αυτό, παρότι τεχνικά ορθό, είναι ερμηνευτικά ελλιπές. Ο αναγνώστης βλέπει δύο καμπύλες που μεταβάλλονται σε πέντε διαδοχικά σημεία, αλλά δεν γνωρίζει τι ακριβώς μετράται. Ο κατακόρυφος άξονας δεν φέρει μονάδες ή περιγραφή, επομένως δεν είναι σαφές αν πρόκειται για χρόνο, απόδοση, ακρίβεια, κόστος ή κάποιο άλλο μέγεθος. Ο οριζόντιος άξονας επίσης δεν εξηγεί τι αναπαριστά: είναι χρονικές περίοδοι, διαφορετικά πειράματα, μέγεθος δείγματος, στάδια διαδικασίας;

Επιπλέον, δεν είναι ξεκάθαρο τι συγκρίνεται. Οι δύο καμπύλες έχουν ασαφή ονόματα ούτε υπόμνημα. Ο παρατηρητής μπορεί να διαπιστώσει ότι η μία γραμμή βρίσκεται σταθερά υψηλότερα από την άλλη, αλλά δεν γνωρίζει ποιο σύστημα αντιστοιχεί σε κάθε γραμμή ούτε ποια είναι η σημασία αυτής της διαφοράς. Η οπτική εντύπωση της υπεροχής δεν μεταφράζεται αυτομάτως σε ουσιαστικό συμπέρασμα.

Προβληματικό είναι και ότι δεν παρέχονται ακριβείς αριθμητικές τιμές. Ο αναγνώστης καλείται να εκτιμήσει κατά προσέγγιση τα μεγέθη από το ύψος των σημείων, κάτι που μπορεί να είναι παραπλανητικό ή ανεπαρκές, ιδίως αν οι διαφορές είναι μικρές αλλά ουσιαστικές. Χωρίς αριθμητική ακρίβεια ή έστω κάποια συμπληρωματική περιγραφή, το γράφημα περιορίζεται σε μια γενική οπτική σύγκριση.

Ας θεωρήσουμε ένα ραβδόγραμμα για να γίνει αυτή η παρατήρηση πιο κατανοητή. Δεν πρέπει να αναμένουμε από τον αναγνώστη να υπολογίσει κατά προσέγγιση τις τιμές συγκρίνοντας τα ύψη. Αν οι ακριβείς αριθμοί έχουν σημασία, πρέπει να αναφέρονται

καθαρά, είτε πάνω στο γράφημα είτε στο κυρίως κείμενο που συνοδεύει την οπτικοποίηση. Η οπτική εντύπωση δεν υποκαθιστά την ποσοτική ακρίβεια.

Ανάλογα, εάν υπάρχει ένα ιδιαίτερα σημαντικό σημείο (όπως μια κρίσιμη περιοχή, ένα κατώφλι ή μια ανωμαλία), αυτό πρέπει να επισημαίνεται οπτικά. Όμως η επισημάνση δεν αρκεί από μόνη της. Η λεζάντα οφείλει να το αναφέρει ρητά, ώστε να κατευθύνει την ερμηνεία και να αποτρέπει παρερμηνείες.

Αντίθετα, στην Εικόνα 6.4 απουσιάζει κάθε επεξηγηματικό πλαίσιο. Δεν γνωρίζουμε αν πρόκειται για πειραματικά δεδομένα, για προσομοίωση ή για πραγματικές μετρήσεις. Δεν γνωρίζουμε αν υπάρχουν σφάλματα μέτρησης, διακυμάνσεις ή στατιστική σημαντικότητα. Το γράφημα λειτουργεί ως ένα απλό σχήμα χωρίς καμία αφήγηση.

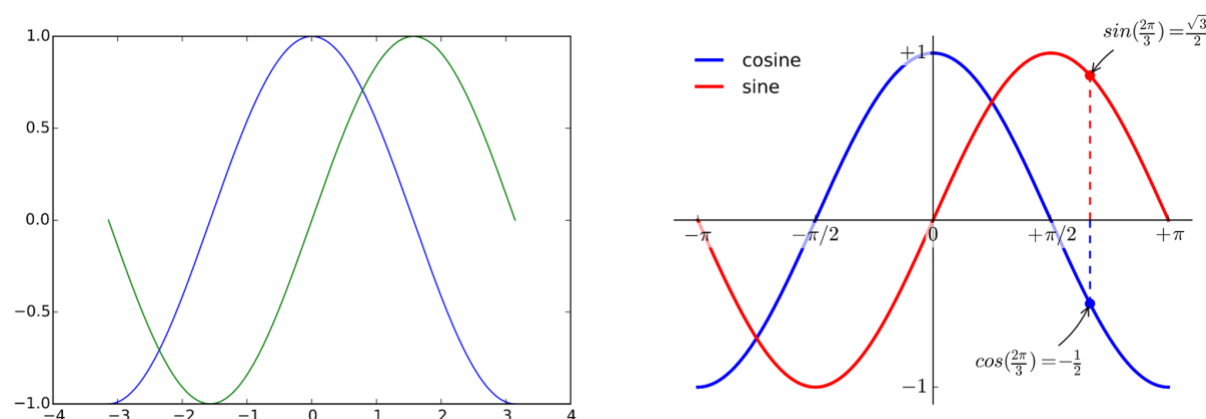
Ο τέταρτος κανόνας, έτσι, αναδεικνύει την οπτικοποίηση ως μία συνεργασία εικόνας και (γραπτού) λόγου. Το γράφημα συλλαμβάνει την προσοχή και οργανώνει την πληροφορία. Η λεζάντα παρέχει το πλαίσιο και την ακρίβεια. Μαζί συγκροτούν κάτι ενιαίο. Αν κάποιο από τα δύο υπολείπεται, η επικοινωνία αποδυναμώνεται.

6.5. Αρχή V: Μην εμπιστεύεστε την προεπιλογή

Κάθε βιβλιοθήκη οπτικοποίησης συνοδεύεται από ένα σύνολο προεπιλεγμένων ρυθμίσεων. Αν ο χρήστης δεν καθορίσει τίποτε, το σύστημα αποφασίζει για το μέγεθος του σχήματος, τη γραμματοσειρά, τα χρώματα, το πλάτος των γραμμών, και τη συνολική αισθητική. Οι ρυθμίσεις αυτές έχουν σχεδιαστεί ώστε να λειτουργούν καλά για την πλειοψηφία των γραφημάτων. Γι' αυτό, όμως, δεν θα έπρεπε να περιμένουμε να είναι οι ιδανικές ρυθμίσεις για έναν συγκεκριμένο τύπο.

Οι προεπιλογές είναι γενικού σκοπού. Δεν γνωρίζουν ποιο είναι το μήνυμα του γραφήματος, ποιο είναι το κοινό, ούτε σε ποιο μέσο θα παρουσιαστεί. Έτσι, ενώ το αποτέλεσμα είναι συνήθως τεχνικά αποδεκτό, σπάνια είναι βέλτιστο. Συχνά μπορούμε να αναγνωρίσουμε την προέλευση ενός γραφήματος (αν έχει παραχθεί σε Matlab, Excel, ή matplotlib) απλώς από τις αισθητικές επιλογές που επιβάλλουν οι προεπιλογές.

Το παράδειγμα της Εικόνας 6.6 παρακάτω είναι χαρακτηριστικό.



Εικόνα 6.517 Προεπιλεγμένες ρυθμίσεις (αριστερά) έναντι ενός βελτιστοποιημένου σχεδιασμού (δεξιά). Από Rougier et al. (2014), Άδεια CC0.

Αριστερά, βλέπουμε το ημίτονο και το συνημίτονο όπως αποδίδονται από το matplotlib με προεπιλεγμένες ρυθμίσεις. Το αποτέλεσμα είναι σωστό. Ωστόσο, δεν υπάρχουν ετικέτες στους άξονες ή στις καμπύλες, τα χρώματα δεν είναι πολύ εμφανή, υπάρχει κενό αριστερά και δεξιά. Δεξιά, έγιναν πιο έντονα τα γράμματα, με μεγαλύτερο πάχος στις γραμμές, χωρίς το περιττό περίγραμμα, με τιμές συναρτήσεων ακτινίων (π), με επισημείωση να εξηγεί τις τριγωνομετρικές συναρτήσεις και να δίνει την τιμή για δεδομένο ακτίνιο στον οριζόντιο άξονα.

Η βασική αρχή είναι ότι καμία προεπιλογή δεν μπορεί να γνωρίζει τον σκοπό του γραφήματος. Οι προεπιλογές αποτελούν μία καλή αφετηρία, αλλά όχι το τελικό αποτέλεσμα.

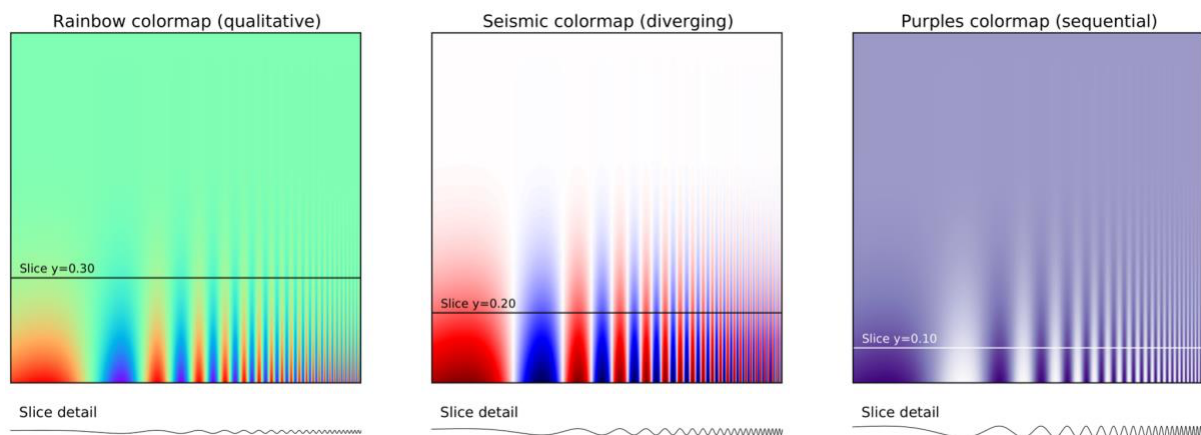
6.6. Αρχή VI: Αποτελεσματική χρήση του χρώματος

Το χρώμα αποτελεί θεμελιώδη διάσταση της ανθρώπινης όρασης και, κατ' επέκταση, σημαντικό εργαλείο στον σχεδιασμό επιστημονικών γραφημάτων. Όπως έχει επισημάνει ο Edward Tufte, το χρώμα μπορεί να είναι ο ισχυρότερος σύμμαχος ή ο χειρότερος εχθρός. Η διαφορά έγκειται στη συνειδητή και στοχευμένη χρήση του.

Η βασική αρχή είναι απλή: χρησιμοποιούμε χρώμα όταν εξυπηρετεί έναν σκοπό. Αν θέλουμε να αναδείξουμε ένα στοιχείο μέσα σε ένα γράφημα, μπορούμε να το χρωματίσουμε έντονα και να διατηρήσουμε τα υπόλοιπα στοιχεία σε γκρι ή μαύρο. Η αντίθεση δημιουργεί οπτική έμφαση. Αν όμως δεν υπάρχει συγκεκριμένος λόγος για τη χρήση χρώματος, οφείλουμε να ρωτήσουμε: «Γιατί αυτό το γράφημα είναι μπλε και όχι μαύρο;». Αν δεν υπάρχει σαφής απάντηση, τότε πιθανότατα το μπλε χρώμα να μην προσθέτει πληροφορία.

Η ίδια αρχή ισχύει και για τις χρωματικές κλίμακες (colormaps). Προεπιλεγμένες κλίμακες, όπως η “rainbow”, δημιουργούν τεχνητά όρια και μη γραμμικές μεταβολές, με αποτέλεσμα να κρύβουν λεπτομέρειες ή να εισάγουν ψευδείς δομές. Γενικώς, οι χρωματικές κλίμακες ταξινομούνται παραδοσιακά σε τρεις βασικές κατηγορίες. Οι **διαδοχικές** (sequential) βασίζονται σε μία βασική απόχρωση με μεταβολή φωτεινότητας και είναι κατάλληλες για ποσοτικά δεδομένα που αυξάνονται από χαμηλές σε υψηλές τιμές. Οι **αποκλίνουσες** (diverging) μεταβαίνουν από ένα χρώμα σε ένα άλλο μέσω ουδέτερου σημείου και χρησιμοποιούνται όταν θέλουμε να αναδείξουμε απόκλιση από μια κεντρική τιμή, όπως ο μέσος όρος ή το μηδέν. Οι **ποιοτικές** (qualitative) περιλαμβάνουν διακριτές και έντονα διαφοροποιημένες αποχρώσεις και ενδείκνυνται για κατηγορικά δεδομένα.

Η Εικόνα 6.6 παρουσιάζει το ίδιο σήμα, με συχνότητα που αυξάνεται προς τα δεξιά και ένταση που αυξάνεται προς τα κάτω, χρησιμοποιώντας τρεις διαφορετικές κλίμακες.



Εικόνα 6.618 Οι τρεις βασικές κατηγορίες χρωματικής κλίμακας. Από αριστερά προς δεξιά, είναι η ποιοτική, η αποκλίνουσα, και η διαδοχική. Από Rougier et al. (2014), Άδεια CC0.

Η κλίμακα ουράνιου τόξου (ποιοτική) και η αποκλίνουσα “seismic” κρύβουν λεπτομέρειες στην περιοχή υψηλής συχνότητας. Αντίθετα, μια διαδοχική κλίμακα (π.χ., μωβ αποχρώσεων) επιτρέπει μία σαφέστερη αντίληψη των λεπτών μεταβολών.

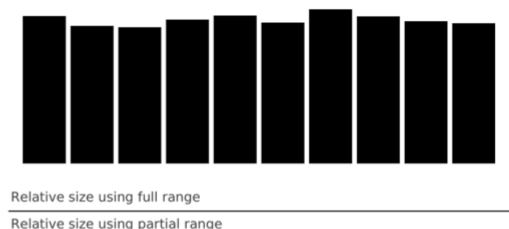
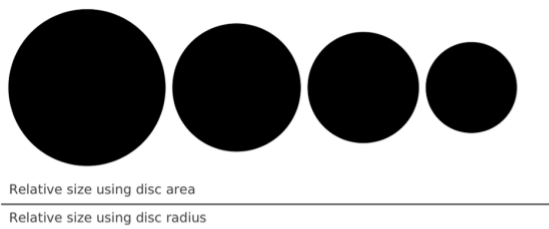
Τα ίδια ακριβώς δεδομένα, λοιπόν, γίνονται περισσότερο ή λιγότερο κατανοητά απλώς λόγω της επιλογής της χρωματικής κλίμακας. Η επιλογή χρωμάτων, επίσης, πρέπει να λαμβάνει υπόψη την αχρωματοψία. Η χρήση πολλών παρόμοιων αποχρώσεων ή συνδυασμών, όπως το κόκκινο–πράσινο, μπορεί να δυσχεραίνει σημαντικά την ανάγνωση για μέρος του πληθυσμού.

Σε γενικές γραμμές, όταν το χρώμα επιλέγεται με βάση τη φύση των δεδομένων και το μήνυμα που θέλουμε να μεταδώσουμε, ενισχύει τη σαφήνεια. Όταν χρησιμοποιείται άκριτα, όμως, μπορεί να εισάγει θόρυβο ή να παραπλανήσει.

6.7. Αρχή VII: Μην παραπλανείτε

Αυτό που διακρίνει ένα επιστημονικό γράφημα από ένα εικαστικό έργο είναι η άμεση σύνδεσή του με κάποια δεδομένα. Η οπτικοποίηση δεν είναι μία ελεύθερη ερμηνεία, αλλά μία αναπαράσταση μετρήσεων, μοντέλων ή αποτελεσμάτων. Αν αυτή η σύνδεση χαλαρώσει, ακόμη και άθελά μας, το μήνυμα είναι λογικό να αλλοιωθεί.

Η αντικειμενική παρουσίαση των δεδομένων δεν είναι πάντα αυτονόητη. Η παραπλάνηση μπορεί να προκύπτει όχι από πρόθεση αλλά από κάποιες προεπιλεγμένες ρυθμίσεις ενός λογισμικού (όπως το matplotlib) ή από φαινομενικά λογικές σχεδιαστικές επιλογές. Η Εικόνα 6.7 παρουσιάζει δύο χαρακτηριστικές τέτοιες περιπτώσεις.



Εικόνα 6.719 Παραπλανητικές αναπαραστάσεις μέσω σχεδιαστικών επιλογών. Αριστερά, τέσσερις τιμές (30, 20, 15, 10) ορίζουν την επιφάνεια (πάνω, σε μαύρο) ή την ακτίνα (κάτω, σε κόκκινο). Δεξιά, οι ίδιες τιμές φαίνονται σε πλήρη κλίμακα (πάνω, σε μαύρο) ή εστιάζοντας σε ένα μικρό διάστημα (κάτω, σε κόκκινο). Από Rougier et al. (2014), Άδεια CC0.

Στο αριστερό μέρος απεικονίζονται τέσσερις τιμές: 30, 20, 15 και 10. Στην πάνω εκδοχή, το μέγεθος των κύκλων αντιστοιχεί στην επιφάνειά τους. Στην κάτω, αντιστοιχεί στην ακτίνα. Η διαφορά είναι καθοριστική. Όταν η τιμή χρησιμοποιείται ως ακτίνα, η επιφάνεια (που είναι το στοιχείο που αντιλαμβάνεται το μάτι) αυξάνεται με το τετράγωνο της ακτίνας. Έτσι, η διαφορά μεταξύ 30 και 10 φαίνεται πολύ μεγαλύτερη (μία οπτική παραμόρφωση) από την πραγματική αναλογία 3:1. Η παρανόηση αυτή είναι συχνή, διότι πολλά εργαλεία ορίζουν την ακτίνα ως προεπιλεγμένη παράμετρο.

Στο δεξί μέρος της εικόνας παρουσιάζεται μια άλλη μορφή παραπλάνησης, η αποκοπή και αλλαγή της κλίμακας. Η ίδια σειρά τιμών απεικονίζεται δύο φορές. Στην επάνω εκδοχή, ο κατακόρυφος άξονας κυμαίνεται από 0 έως 100, ενώ στην κάτω από 80 έως 100. Πάνω, οι διαφορές φαίνονται μικρές· κάτω, όμως, είναι εντυπωσιακές. Παρότι οι τιμές είναι ίδιες, η οπτική εντύπωση διαφέρει ριζικά. Ακόμη και αν η κλίμακα δηλώνεται ρητά, οι κόκκινες ράβδοι θα συνέχιζαν να ενισχύουν την αίσθηση κάποιας μεγάλης διαφοράς μεταξύ τους.

Αυτά τα παραδείγματα αναδεικνύουν δύο μορφές σφάλματος. Η πρώτη προκύπτει από ακούσιες επιλογές που κληροδοτούν τα εργαλεία μας μέσω προεπιλεγμένων ρυθμίσεων. Για παράδειγμα, η αυτόματη κλιμάκωση αξόνων, η επιλογή συγκεκριμένων χρωματικών παλετών ή ο τρόπος με τον οποίο αποδίδεται το μέγεθος σημείων σε ένα γράφημα μπορεί να είναι τεχνικά ορθός, αλλά να οδηγεί σε οπτικά παραπλανητικές εντυπώσεις.

Η δεύτερη μορφή προκύπτει από ρητές αλλά προβληματικές αποφάσεις, όπως η χρήση τρισδιάστατων γραφημάτων ή διαγραμμάτων πίτας για ποσοτική σύγκριση. Στις περιπτώσεις αυτές, η γεωμετρική απεικόνιση δεν αντιστοιχεί άμεσα στον τρόπο με τον οποίο ο άνθρωπος αντιλαμβάνεται τις διαφορές μεγεθών, με αποτέλεσμα να υπερτονίζονται ή να υποβαθμίζονται σχέσεις μεταξύ τιμών. Παρότι οι επιλογές αυτές γίνονται συνειδητά, συχνά βασίζονται σε αισθητικά κριτήρια και όχι σε κριτήρια ακρίβειας.

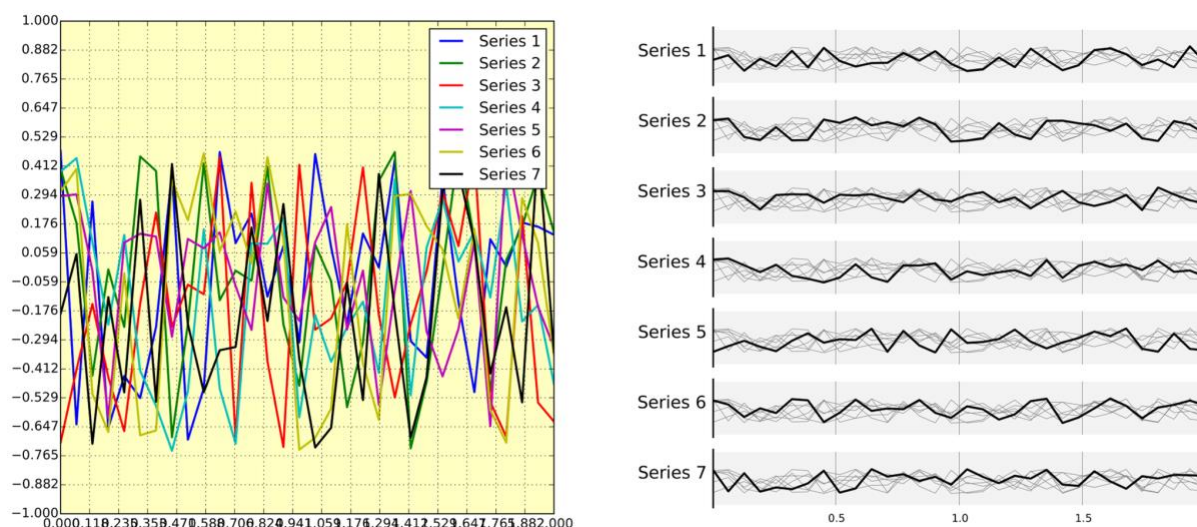
Ο γενικός κανόνας είναι απλός: *επιλέγουμε τον απλούστερο τύπο γραφήματος που μπορεί να μεταδώσει το μήνυμα και χρησιμοποιούμε πλήρες εύρος τιμών όταν αυτό είναι ουσιώδες για την ερμηνεία.*

Η ευθύνη για τη σαφήνεια και την ειλικρίνεια της οπτικοποίησης ανήκει πάντοτε στον δημιουργό. Αν αναλάβουμε την ευθύνη αυτή σοβαρά, μία πρακτική συμβουλή είναι να ζητάμε από συναδέλφους να ερμηνεύσουν τα γραφήματά μας. Αν η αυθόρμητη ερμηνεία τους διαφέρει από το επιδιωκόμενο μήνυμα, τότε το γράφημα χρειάζεται αναθεώρηση.

6.8. Αρχή VIII: Μείωση του οπτικού θορύβου

Ο οπτικός θόρυβος αναφέρεται σε κάθε οπτικό στοιχείο που δεν βελτιώνει το μήνυμα ενός γραφήματος (περιττό) ή που το αποδυναμώνει (π.χ., προκαλεί σύγχυση). Σύμφωνα με τον Edward Tufte, ότι δε μεταφέρει νέα πληροφορία είναι *chartjunk* και θα πρέπει να αφαιρείται.

Θόρυβος μπορεί να προκύψει από την υπερβολική χρήση χρωμάτων, περιττές ετικέτες, από κάποιο βαρύ φόντο, άσκοπες γραμμές πλέγματος, από υπερβολική ακρίβεια δεκαδικών, από διακοσμητικά εφέ ή και από την επικάλυψη στοιχείων που δυσχεραίνουν την ανάγνωση. Η υπερφόρτωση δεν αυξάνει την πληροφορία αλλά τον θόρυβο.



Εικόνα 6.820 Παράδειγμα υπερφόρτωσης (αριστερά) και καθαρής σχεδίασης (δεξιά). Από Rougier et al. (2014), Άδεια CC0.

Στο αριστερό μέρος της Εικόνας 6.8 παρουσιάζονται επτά σειρές δεδομένων με τρόπο που καθιστά σχεδόν αδύνατη τη σύγκριση. Οι καμπύλες επικαλύπτονται, τα αυτόματα επιλεγμένα χρώματα δεν βοηθούν στη διάκριση, το υπόμνημα καλύπτει μέρος του γραφήματος, τα ticks είναι υπερβολικά και αλληλοεπικαλυπτόμενα, ενώ το πλέγμα παραπέμπει σε τιμές που είναι μάλλον υπερβολικά αναλυτικές. Η συνολική εντύπωση είναι σύγχυση.

Στο δεξί μέρος, το ίδιο περιεχόμενο επανασχεδιάζεται στον ίδιο διαθέσιμο χώρο. Κάθε σειρά παρουσιάζεται σε ξεχωριστό υπογράφημα, ενώ οι υπόλοιπες εμφανίζονται διακριτικά στο υπόβαθρο για σύγκριση. Οι ετικέτες τοποθετούνται άμεσα δίπλα σε κάθε σειρά, εξαλείφοντας την ανάγκη υπομνήματος και πολλών χρωμάτων. Τα ticks μειώνονται στα απολύτως απαραίτητα (τρία), και μια λεπτή διακριτική γραμμή τα συνδέει οπτικά μεταξύ των επιμέρους υπογραφημάτων. Οι τιμές του κάθετου άξονα αφαιρούνται πλήρως, ενώ ένα γκρι πλαίσιο υποδηλώνει το εύρος $[-1, +1]$, το οποίο θα πρέπει να διευκρινίζεται στη λεζάντα.

Η διαφορά μεταξύ των δύο εκδοχών δεν αφορά τα δεδομένα αλλά τον σχεδιασμό. Το ίδιο περιεχόμενο μπορεί να μην είναι δυσανάγνωστο και να γίνει σαφές, απλώς αφαιρώντας περιττά στοιχεία και ενισχύοντας τα ουσιαστικά, με βάση την εφαρμογή, το μήνυμα, το κοινό.

6.9. Αρχή IX: Το μήνυμα υπερισχύει της αισθητικής

Τα επιστημονικά γραφήματα δεν είναι απλώς εικόνες που συνοδεύουν ένα άρθρο ή μια εργασία. Αποτελούν βασικό εργαλείο ανάλυσης και τεκμηρίωσης, ιδιαίτερα στην Πληροφορική, όπου συγκρίνουμε αλγορίθμους, αξιολογούμε μοντέλα και μελετούμε πειραματικά αποτελέσματα. Σε βάθος χρόνου, κάθε υποπεδίο (από τη μηχανική μάθηση έως τα συστήματα και τα δίκτυα) έχει διαμορφώσει συγκεκριμένες συμβάσεις οπτικοποίησης. Οι συμβάσεις αυτές διευκολύνουν τη σύγκριση αποτελεσμάτων και συμβάλλουν στον έλεγχο της ορθότητας των αποτελεσμάτων. Ένα απρόσμενο γράφημα, για παράδειγμα, μπορεί να αποκαλύψει σφάλμα στον κώδικα, στα δεδομένα ή στις υποθέσεις κάποιας μελέτης.

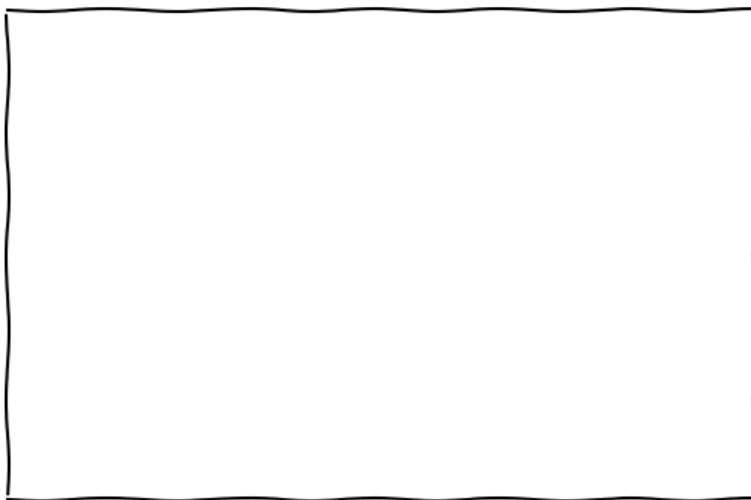
Ωστόσο, συχνά ερχόμαστε αντιμέτωποι με την ανάγκη σχεδιασμού ενός νέου τύπου γραφήματος, για τον οποίο δεν υπάρχει καθιερωμένο πρότυπο. Στην επιστήμη των δεδομένων, η αναγνωσιμότητα και η σαφήνεια του μηνύματος είναι ο πρωταρχικός στόχος. Ένα γράφημα μπορεί να είναι ελκυστικό, αλλά αν θολώνει το μήνυμα, αποτυγχάνει στον επιστημονικό του ρόλο. Από την άλλη πλευρά, τι κάνουμε αν θέλουμε να δώσουμε ένα μήνυμα που δεν υπάρχει μέσα σε δεδομένα (π.χ., για διδακτικούς σκοπούς ή για αφηγηματική απεικόνιση, όχι για παρουσίαση πραγματικών δεδομένων);

Ας το δούμε με ένα παράδειγμα, προσπαθώντας να δείξουμε τι συμβαίνει στην κατανόηση ενός κειμένου καθώς υιοθετούμε όλο και περισσότερο τεχνικές επεξεργασίας φυσικής γλώσσας. Θα χρησιμοποιήσουμε το στυλ `xkcd` της βιβλιοθήκης `matplotlib`, για να έχουμε ένα αποτέλεσμα που θα είναι χιουμοριστικό και αφηγηματικό.

Αρχικά, ας δούμε πώς φτιάχνεται το πλαίσιο σε αυτό το στυλ:

```
import matplotlib.pyplot as plt
import numpy as np

with plt.xkcd():
    fig = plt.figure() # build the figure
    ax = fig.add_axes((0.1, 0.2, 0.8, 0.7))
    ax.set_xticks([]) # show no numbers on the axes
    ax.set_yticks([])
    ax.set_ylim([-30, 10]) # limit the vertical axis
```



Εικόνα 6.921 Πλαίσιο που δίνει την εντύπωση σκίτσου μέσω του στυλ `xkcd`.

Έχει φτιαχτεί το πλαίσιο που δίνει την εντύπωση ενός σκίτσου. Το στυλ `xkcd` προσομοιώνει έναν χειρόγραφο και πρόχειρο σχεδιασμό τύπου κόμικ, με τις γραμμές να αποκτούν ελαφριά «τρεμάμενη» μορφή και τη συνολική εντύπωση να απομακρύνεται από την αυστηρή επιστημονική αισθητική.

Ας προσθέσουμε τώρα τον κώδικα που θα χτίσει το μήνυμα του σκίτσου:

```
with plt.xkcd():
    fig = plt.figure(figsize=(10, 6))
    ax = fig.add_axes((0.1, 0.2, 0.8, 0.7))
    ax.spines[['top', 'right']].set_visible(False)
    ax.set_xticks([])
    ax.set_yticks([])
    ax.set_ylim([-30, 15])

    # the data: The "Sense of Meaning" Curve
    x = np.linspace(0, 10, 100)
    y = 5 - 0.5 * x**2 # A dramatic drop in "Human Intuition"

    # the main drawing
    ax.plot(x, y, color='teal', lw=3)

    # Labeling the stages
    ax.annotate('Reading "Hamlet"\n(Pure Vibes)',
                xy=(0, 5), xytext=(1, 10),
                arrowprops=dict(arrowstyle='->', color='black'))

    ax.annotate('Tokenization:\n"To", "be", "or"...',
                xy=(3.5, -1), xytext=(4.5, 4),
                arrowprops=dict(arrowstyle='->', color='black'))

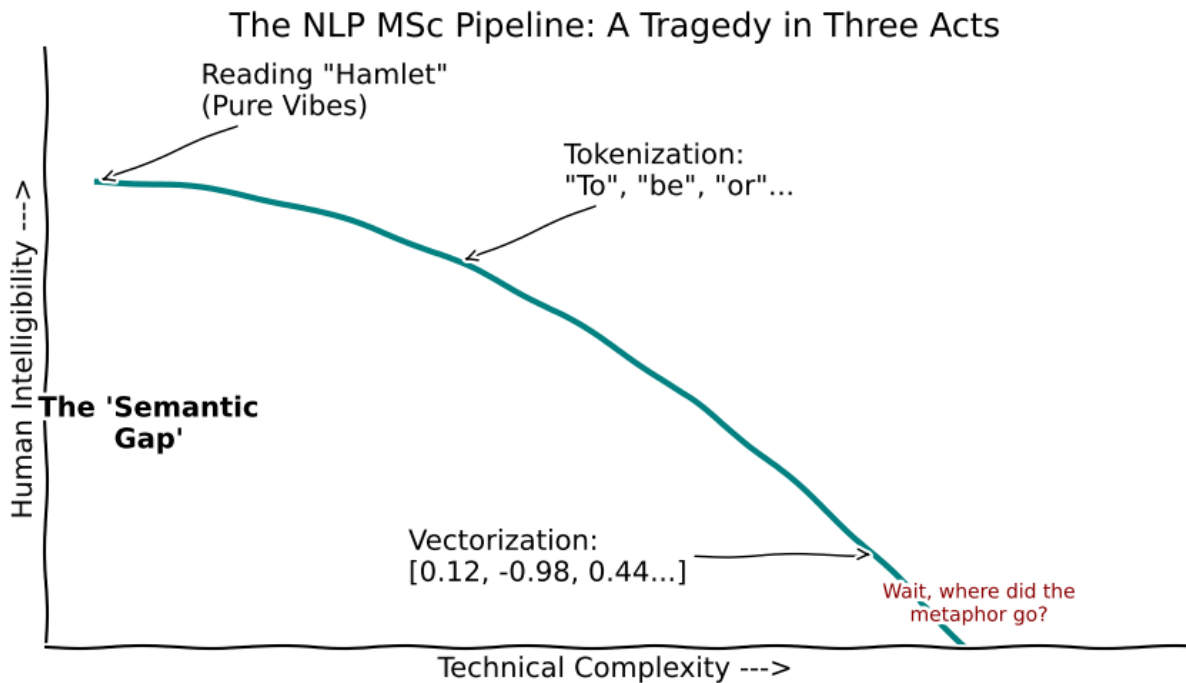
    ax.annotate('Vectorization:\n[0.12, -0.98, 0.44...]',
                xy=(7.5, -23), xytext=(3, -25),
                arrowprops=dict(arrowstyle='->', color='black'))

    # Adding some "Humanities Student" commentary
    plt.text(8.5, -28, "Wait, where did the\nmetaphor go?",
             ha='center', fontsize=10, color='darkred')

    plt.text(0.5, -15, "The 'Semantic\nGap'",
             ha='center', fontweight='bold')

    plt.title("The NLP MSc Pipeline: A Tragedy in Three Acts")
    plt.xlabel("Technical Complexity --->")
    plt.ylabel("Human Intelligibility --->")

    plt.show()
```



Εικόνα 6.1022 Καμπύλη που βασίζεται σε μία πολυωνυμική συνάρτηση για να αποδώσει πώς μειώνεται η ικανότητα του ανθρώπου να κατανοεί καθώς αυξάνεται η τεχνική πολυπλοκότητα. Τρεις επισημειώσεις δείχνουν παραδείγματα της επεξεργασίας φυσικής γλώσσας, από την ανάγνωση του Άμλετ, στο επεξεργασμένο κείμενο, σε διανυσματικές αναπαραστάσεις.

Το γράφημα αυτό αφηγείται μια ιστορία: καθώς αυξάνεται η τεχνική πολυπλοκότητα, μειώνεται η ικανότητα του ανθρώπου να κατανοεί. Είναι ευρηματικό και επιτηδευμένα παιγνιώδες. Οι άξονες δεν έχουν μονάδες, γιατί τις αφαιρέσαμε με την `set_xticks([])`. Η καμπύλη δεν βασίζεται σε πραγματικά δεδομένα, αλλά σε μία πολυωνυμική συνάρτηση. Επίσης, η οπτική παραμόρφωση του **xkcd** στυλ εισάγει μία εσκεμμένη «ακαταστασία». Παρότι δεν θα το χρησιμοποιούσαμε για επιστημονική τεκμηρίωση, το γράφημα αυτό λειτουργεί εξαιρετικά για διδακτικούς σκοπούς, γιατί αποδίδει το μήνυμα σε μία εικόνα.

6.10. Αρχή X: Επιλογή του κατάλληλου εργαλείου

Η επιλογή του εργαλείου οπτικοποίησης αποτελεί άλλη μία απόφαση στη διαδικασία ανάλυσης και παρουσίασης δεδομένων. Η απόφαση αυτή πρέπει να καθοδηγείται από το είδος των δεδομένων, το κοινό στο οποίο απευθύνεται η οπτικοποίηση και το μέσο δημοσίευσης. Παρότι δεν υπάρχει «καλύτερο» εργαλείο γενικά, υπάρχει καταλληλότερο εργαλείο για κάθε συγκεκριμένο σκοπό.

Πρακτικά, το εργαλείο που χρησιμοποιείται για την ανάλυση δεν είναι απαραίτητο να ταυτίζεται με εκείνο που χρησιμοποιείται για την τελική οπτικοποίηση. Τα δεδομένα μπορούν να εξαχθούν και να αξιοποιηθούν σε διαφορετικό περιβάλλον, εφόσον αυτό προσφέρει καταλληλότερες δυνατότητες σχεδιασμού. Στο πλαίσιο της επιστήμης δεδομένων, δύο βιβλιοθήκες ξεχωρίζουν για την ευρεία χρήση και την ευελιξία τους: **Matplotlib** και **Seaborn**.

Στην πράξη, όπως θα δούμε στο επόμενο κεφάλαιο, οι δύο αυτές βιβλιοθήκες λειτουργούν συμπληρωματικά. Η Seaborn επιτρέπει γρήγορη και καθαρή στατιστική απεικόνιση, ενώ η Matplotlib παρέχει λεπτομερή έλεγχο όταν απαιτείται προσαρμογή σε επίπεδο δημοσίευσης.

Πέραν αυτών, υπάρχουν και άλλα εργαλεία που καλύπτουν ειδικές ανάγκες. Η R προσφέρει ισχυρές στατιστικές και γραφικές δυνατότητες. Το Inkscape και το GIMP είναι χρήσιμα για διανυσματική επεξεργασία και τελική αισθητική βελτίωση. Τα TikZ/PGF ενσωματώνονται σε έγγραφα LaTeX για τυπογραφική συνέπεια. Η D3.js επιτρέπει διαδραστικές οπτικοποιήσεις στον ιστό. Το Cytoscape και το Circos εξειδικεύονται σε πολύπλοκα δίκτυα και σχέσεις πολλαπλών επιπέδων.

Για τους περισσότερους, οι βιβλιοθήκες Matplotlib και Seaborn καλύπτουν το μεγαλύτερο μέρος των αναγκών οπτικοποίησης στην Επιστήμη Δεδομένων. Αυτός είναι ο λόγος που τις αναλύουμε περισσότερο στο επόμενο κεφάλαιο. Η ουσιαστική πρόκληση, ωστόσο, δεν είναι τόσο η εκμάθηση πολλών εργαλείων, όσο η κατανόηση των αρχών της ορθής οπτικοποίησης και η ικανότητα επιλογής του κατάλληλου εργαλείου για τον εκάστοτε στόχο.

7. Βιβλιοθήκες οπτικοποίησης στην Python

Η Python διαθέτει ένα πλούσιο οικοσύστημα βιβλιοθηκών για την οπτικοποίηση δεδομένων. Οι βιβλιοθήκες αυτές διαφέρουν ως προς τον βαθμό ελέγχου που παρέχουν, το είδος των γραφημάτων που υποστηρίζουν και το επίπεδο διαδραστικότητας που προσφέρουν. Ορισμένες είναι κατάλληλες για επιστημονικές δημοσιεύσεις και στατικές απεικονίσεις υψηλής ακρίβειας, ενώ άλλες έχουν σχεδιαστεί για διαδραστικές εφαρμογές ή για ειδικούς τύπους δεδομένων, όπως τα γεωχωρικά δεδομένα.

Στο κεφάλαιο αυτό παρουσιάζονται τέσσερις βιβλιοθήκες που χρησιμοποιούνται συχνά στην επιστήμη δεδομένων: **Matplotlib**, **Seaborn**, **Folium** και **Plotly**. Κάθε μία καλύπτει διαφορετικές ανάγκες οπτικοποίησης και συχνά χρησιμοποιούνται συμπληρωματικά.

7.1. Βασική οπτικοποίηση μέσω Matplotlib

Η **Matplotlib** αποτελεί τη θεμελιώδη βιβλιοθήκη οπτικοποίησης σε Python και χρησιμοποιείται εκτενώς στην επιστημονική ανάλυση δεδομένων. Αυτό συμβαίνει γιατί η φιλοσοφία της βιβλιοθήκης βασίζεται στην αναπαράσταση του γραφήματος ως ένα σύνολο αντικειμένων (figure, axes, artists), επιτρέποντας την προσαρμογή της εμφάνισης και της δομής του σχήματος που θέλουμε να κατασκευάσουμε, δίνοντας σχεδόν τον πλήρη έλεγχο σε όλα τα στοιχεία του γραφήματος: από τη θέση των αξόνων και την εμφάνιση των ticks μέχρι τις γραμματοσειρές, τις παλέτες χρωμάτων και τις επισημάνσεις (annotations). Επίσης, παρόλο που υποστηρίζει κυρίως δισδιάστατα (2D) γραφήματα, όπως γραμμικά διαγράμματα, ραβδογράμματα, ιστογράμματα και διαγράμματα διασποράς, παρέχει βασική υποστήριξη και για τρισδιάστατες (3D) απεικονίσεις.

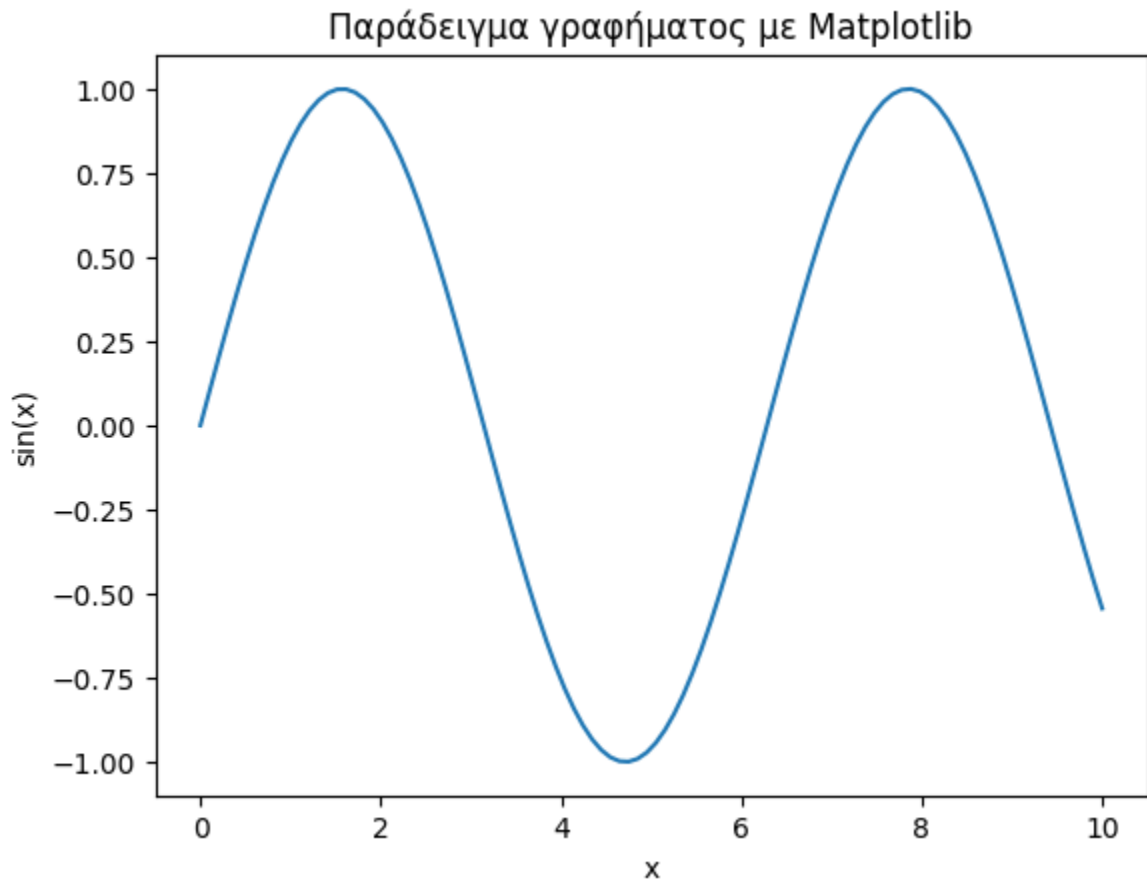
Η βιβλιοθήκη Matplotlib, όπως επισημαίνει και το όνομά της, αναπτύχθηκε με στόχο να προσφέρει στην Python δυνατότητες γραφημάτων αντίστοιχες με εκείνες του Matlab, αλλά με μεγαλύτερη ευελιξία και δυνατότητα προγραμματιστικού ελέγχου.

Ένα παράδειγμα απλού γραφήματος ακολουθεί παρακάτω:

```
import matplotlib.pyplot as plt
import numpy as np

x = np.linspace(0, 10, 100)
y = np.sin(x)

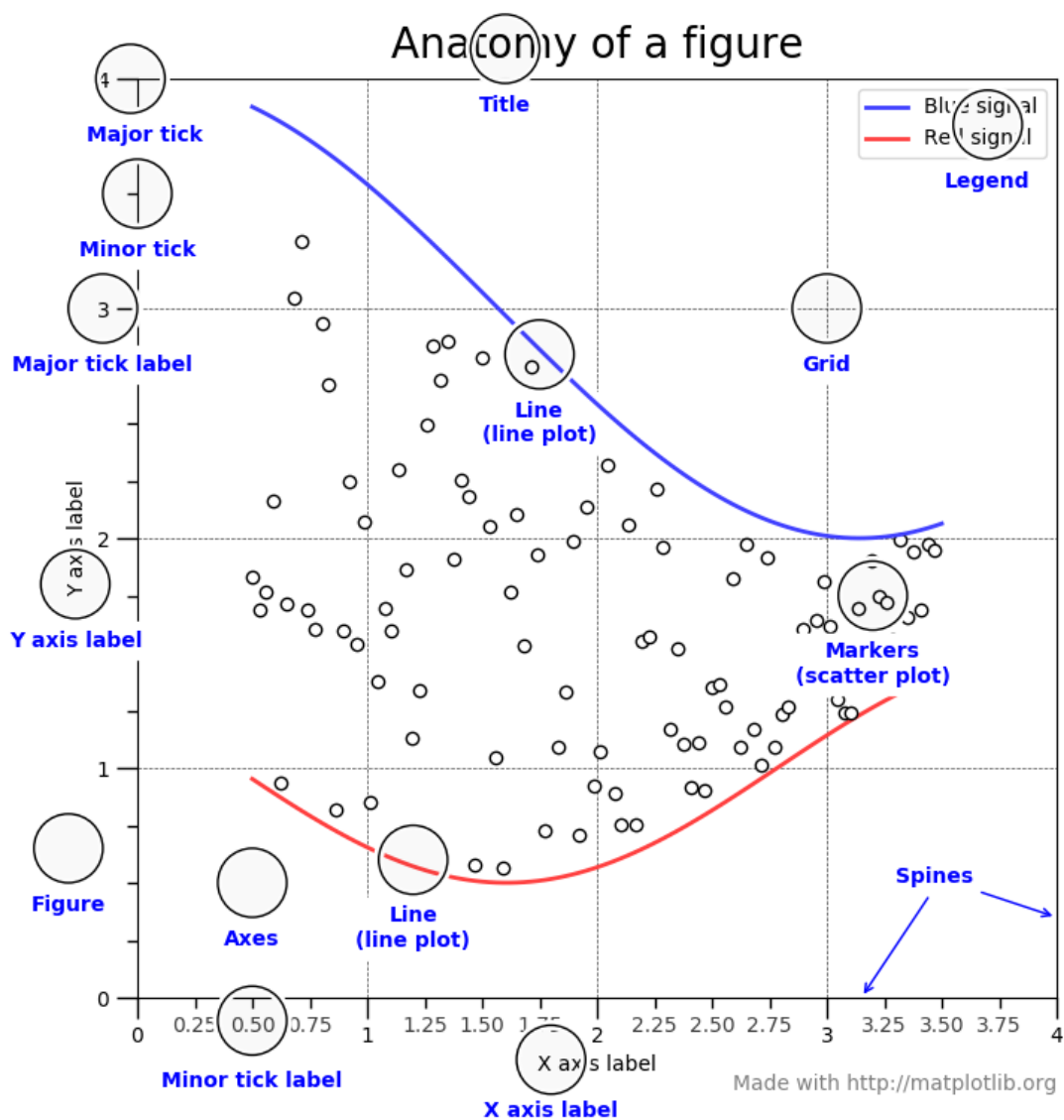
plt.plot(x, y)
plt.xlabel("x")
plt.ylabel("sin(x)")
plt.title("Παράδειγμα γραφήματος με Matplotlib")
plt.show()
```



Εικόνα 7.123 Το γράφημα μιας ημιτονοειδούς συνάρτησης για 100 σημεία μέσω της βιβλιοθήκης Matplotlib.

7.1.1. Ανατομία ενός γραφήματος

Η Matplotlib οργανώνει τα γραφήματα σε μια ιεραρχική δομή αντικειμένων. Ο συνολικός καμβάς ονομάζεται **Figure** και περιέχει ένα ή περισσότερα γραφήματα. Η περιοχή στην οποία σχεδιάζεται το γράφημα ονομάζεται **Axes** ενώ οι άξονες x και y είναι τα **Axis**. Τα επιμέρους οπτικά στοιχεία (γραμμές, σημεία, κείμενα κ.λπ.), τότε, καλούνται **Artists**.



Εικόνα 7.224 Ανατομία ενός matplotlib Figure. Πηγή: <https://matplotlib.org/3.0.2/gallery/showcase/anatomy.html>

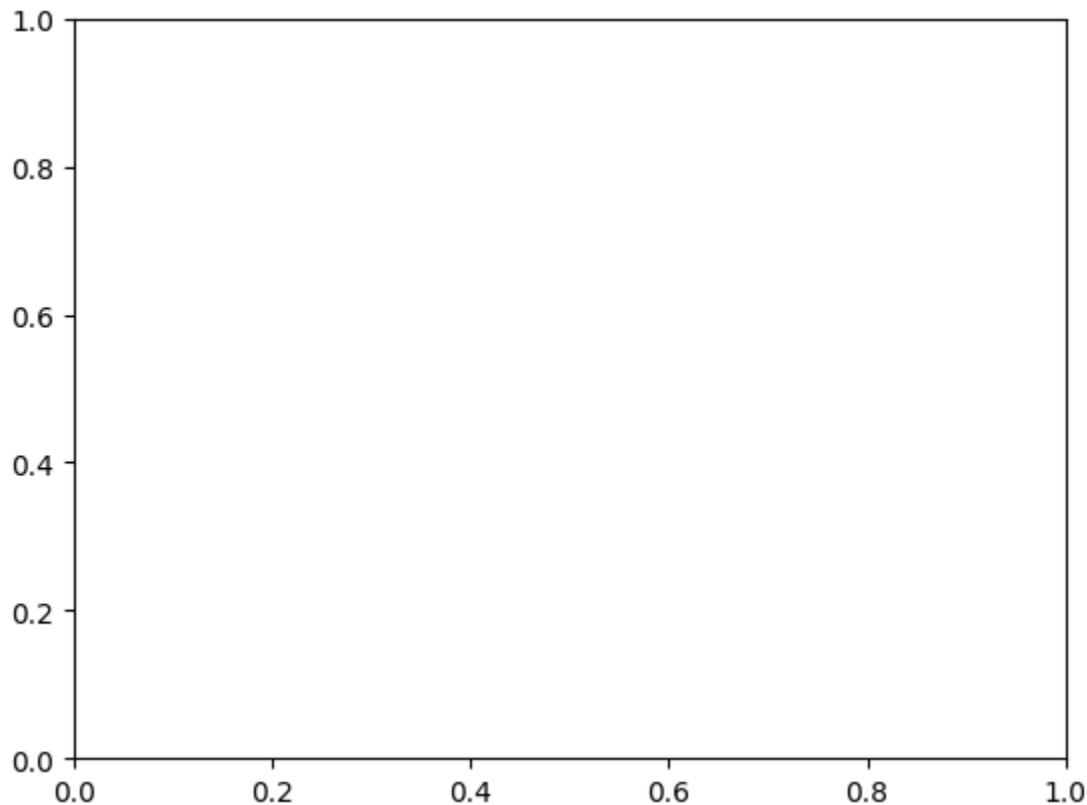
Η βιβλιοθήκη οργανώνει τα γραφήματα ως μια ιεραρχία αντικειμένων. Η παραπάνω εικόνα παρουσιάζει τα βασικά συστατικά ενός γραφήματος στη Matplotlib, που συνεργάζονται για να δώσουν το τελικό αποτέλεσμα της οπτικοποίησης.

Το **Figure** είναι το αντικείμενο που περιέχει όλα τα υπόλοιπα στοιχεία και καθορίζει το συνολικό μέγεθος της εικόνας και τον χώρο μέσα στον οποίο θα τοποθετηθούν τα επιμέρους γραφήματα. Δημιουργείται, για παράδειγμα, ως εξής:

```
fig = plt.figure()
```

ή

```
fig, ax = plt.subplots()
```



Εικόνα 7.325 Κενό γράφημα (Figure) μέσω της βιβλιοθήκης Matplotlib.

Η περιοχή μέσα στο figure όπου σχεδιάζεται το γράφημα (**Axes**) περιλαμβάνει τους άξονες, τις καμπύλες, τις επισημάνσεις και τα περισσότερα οπτικά στοιχεία. Ένα γράφημα μπορεί να έχει περισσότερα από ένα αντικείμενο axes (π.χ., subplots). Στην περίπτωση αυτή, η μεταβλητή **ax** αναφέρεται στο συγκεκριμένο γράφημα μέσα στο figure.

Κάθε αντικείμενο axes περιλαμβάνει δύο βασικούς άξονες, τον οριζόντιο (**x-axis**) και τον κάθετο (**y-axis**). Με τη σειρά τους αυτοί οι άξονες περιλαμβάνουν διαβαθμίσεις (ticks), αριθμητικές ενδείξεις (tick labels) και ετικέτες αξόνων (axis labels). Οι άξονες καθορίζουν την κλίμακα και την ερμηνεία των δεδομένων. Αξίζει να σημειωθεί ότι η Matplotlib δημιουργεί αυτόματα ticks, αλλά μπορούμε να τα τροποποιήσουμε ή να τα αφαιρέσουμε:

```
ax.set_xticks([0, 5, 10])
```

Επίσης, στην εικόνα φαίνονται τα περιγράμματα του γραφήματος (**spines**) που ορίζουν το πλαίσιο των αξόνων. Συνήθως εμφανίζονται τέσσερα: πάνω, κάτω, αριστερά και δεξιά. Ωστόσο, αν αφαιρεθούν ορισμένα μπορεί να δοθεί μία πιο καθαρή αισθητική:

```
ax.spines["top"].set_visible(False)
ax.spines["right"].set_visible(False)
```

Ο **τίτλος** του γραφήματος παρέχει μία συνοπτική περιγραφή του περιεχομένου:

```
ax.set_title("Παράδειγμα γραφήματος")
```

Η λεζάντα (**legend**) εξηγεί τις σειρές δεδομένων που εμφανίζονται στο γράφημα:

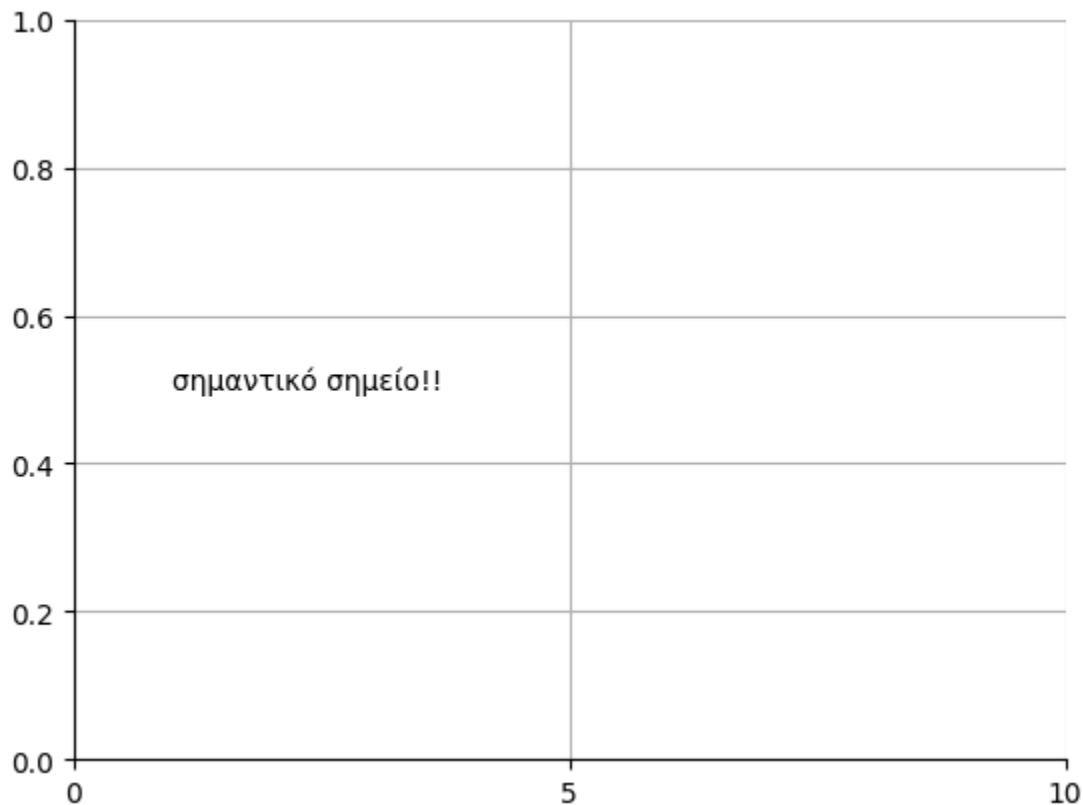
```
ax.legend()
```

Το πλέγμα γραμμών (**grid**) βοηθά στην εκτίμηση των τιμών πάνω στο γράφημα.

```
ax.grid(True)
```

Τέλος, μπορούμε να επισημάνουμε συγκεκριμένα σημεία ενδιαφέροντος στο γράφημα μέσω επισημάνσεων (**annotations**):

```
ax.annotate("σημαντικό σημείο!!", xy=(0,0), xytext=(1, 0.5))
```



Εικόνα 7.426 Επισημείωση ενός σημείου στο γράφημα μέσω της βιβλιοθήκης Matplotlib.

7.1.2. Οπτικοποίηση με αντικειμενοστραφή προγραμματισμό

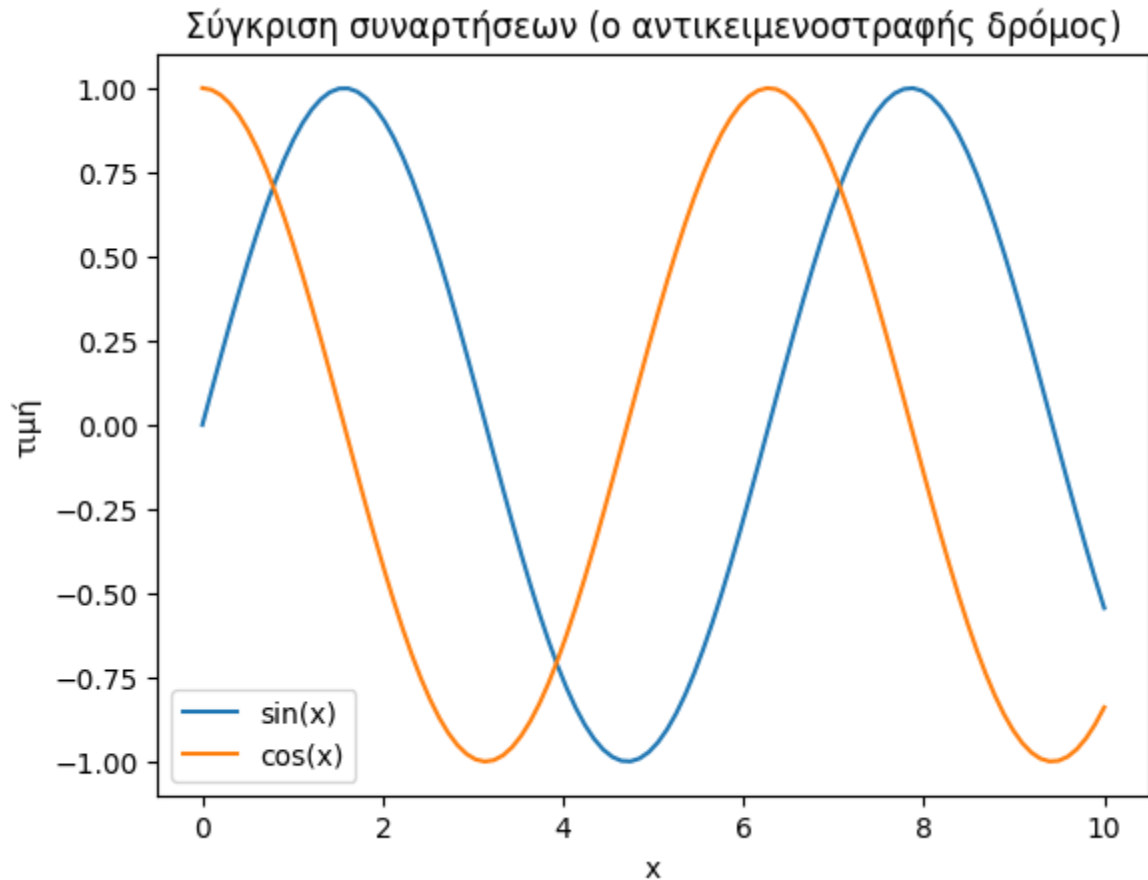
Η Matplotlib προσφέρει δύο βασικούς τρόπους δημιουργίας γραφημάτων. Ο πρώτος (pyplot) θυμίζει MATLAB, όπου οι εντολές εφαρμόζονται σε κάποιο γράφημα. Ο δεύτερος αφορά την αντικειμενοστραφή (object-oriented) προσέγγιση, όπου το γράφημα αντιμετωπίζεται ως ένα σύνολο αντικειμένων στα οποία μπορούμε να καλέσουμε μεθόδους. Δημιουργούμε **πρώτα τα αντικείμενα** και κατόπιν τα χειριζόμαστε. Αυτό προσφέρει σαφή κώδικα και καλύτερο έλεγχο όταν έχουμε πιο πολλά (ταυτόχρονα) ή σύνθετα γραφήματα.

Το παρακάτω παράδειγμα δείχνει την αντικειμενοστραφή προσέγγιση:

```
import numpy as np
import matplotlib.pyplot as plt

x = np.linspace(0, 10, 100)
fig, ax = plt.subplots()
```

```
ax.plot(x, np.sin(x), label="sin(x)")
ax.plot(x, np.cos(x), label="cos(x)")
ax.set_xlabel("x")
ax.set_ylabel("τιμή")
ax.set_title("Σύγκριση συναρτήσεων (ο αντικειμενοστραφής δρόμος)")
ax.legend()
plt.show()
```

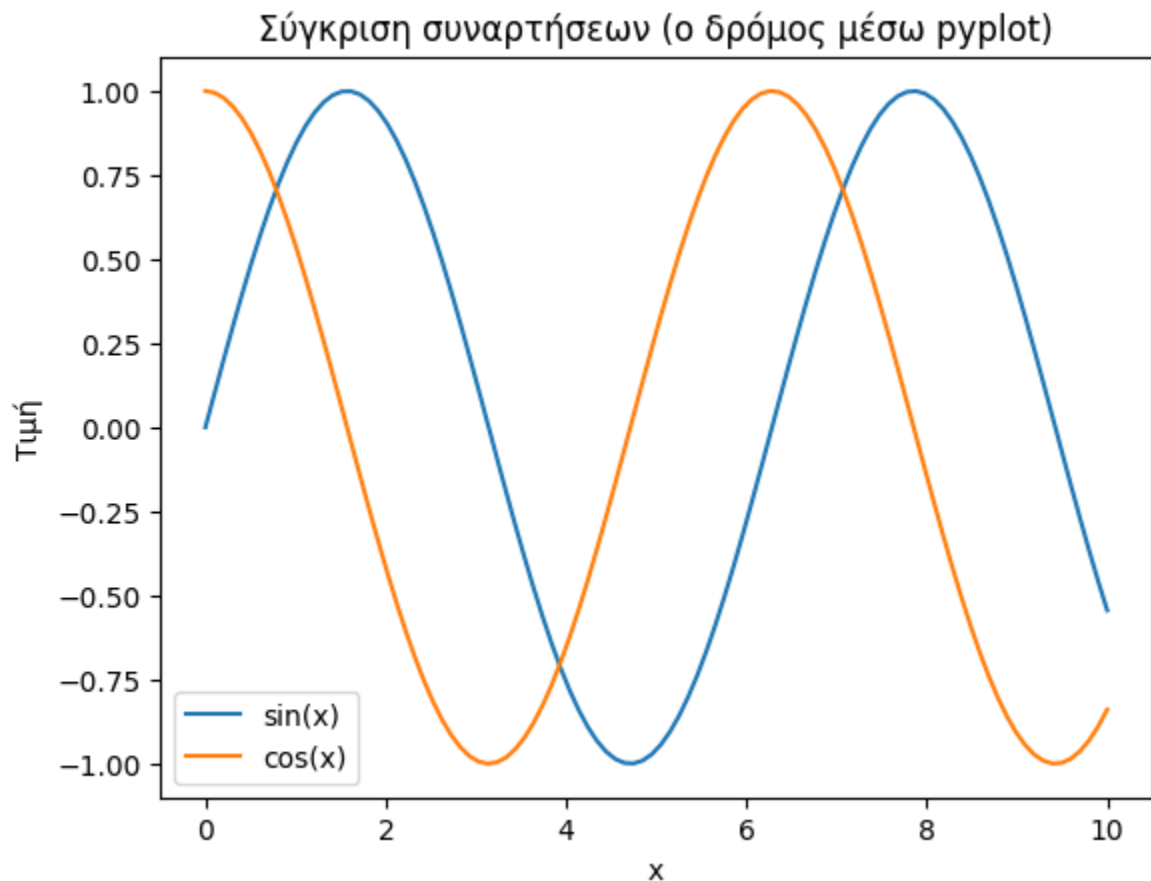


Εικόνα 7.527 Συνάρτηση ημιτόνου (μπλε) και συνημιτόνου (πορτοκαλί) σε ένα γράφημα που φτιάχτηκε με την αντικειμενοστραφή προσέγγιση προγραμματισμού.

Το fig είναι ένα αντικείμενο Figure (το συνολικό πλαίσιο) και το ax είναι ένα αντικείμενο Axes, δηλαδή το συγκεκριμένο γράφημα. Μέθοδοι όπως plot, set_title, set_xlabel και set_ylabel καλούνται απευθείας στο αντικείμενο ax.

Το ίδιο γράφημα με τον pyplot τρόπο θα γινόταν ως εξής:

```
x = np.linspace(0, 10, 100)
plt.plot(x, np.sin(x), label="sin(x)")
plt.plot(x, np.cos(x), label="cos(x)")
plt.title("Σύγκριση συναρτήσεων (ο δρόμος μέσω pyplot)")
plt.xlabel("x")
plt.ylabel("Τιμή")
plt.legend()
plt.show()
```



Εικόνα 7.628 Το ίδιο αποτέλεσμα με την Εικόνα 7.5 αλλά μέσω του pyplot τρόπου.

Παρότι απλούστερη για γρήγορα γραφήματα, η pyplot προσέγγιση γίνεται δυσκολότερη στη διαχείριση όταν υπάρχουν πολλά υπογραφήματα ή πιο σύνθετες διατάξεις.

7.1.3. Πολλαπλά υπογραφήματα

Η Matplotlib επιτρέπει την τοποθέτηση πολλών γραφημάτων στο ίδιο σχήμα, κάτι ιδιαίτερα χρήσιμο για συγκρίσεις (π.χ., για επιστημονικές δημοσιεύσεις).

Ακολουθεί ένα τέτοιο παράδειγμα, όπου ακολουθούμε την αντικειμενοστραφή προσέγγιση:

```
import numpy as np
import matplotlib.pyplot as plt

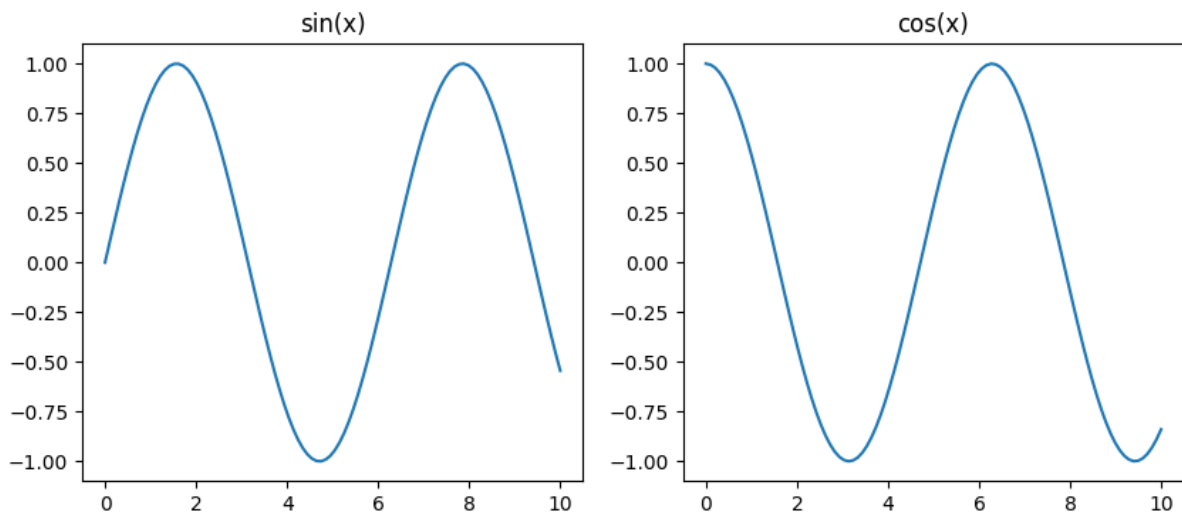
x = np.linspace(0, 10, 100)

fig, axes = plt.subplots(1, 2, figsize=(10,4))

axes[0].plot(x, np.sin(x))
axes[0].set_title("sin(x)")

axes[1].plot(x, np.cos(x))
axes[1].set_title("cos(x)")

plt.show()
```



Εικόνα 7.729 Δύο υπογραφήματα, ως αντικείμενα Axes, για την συνάρτηση ημιτόνου (αριστερά) και συνημιτόνου (δεξιά).

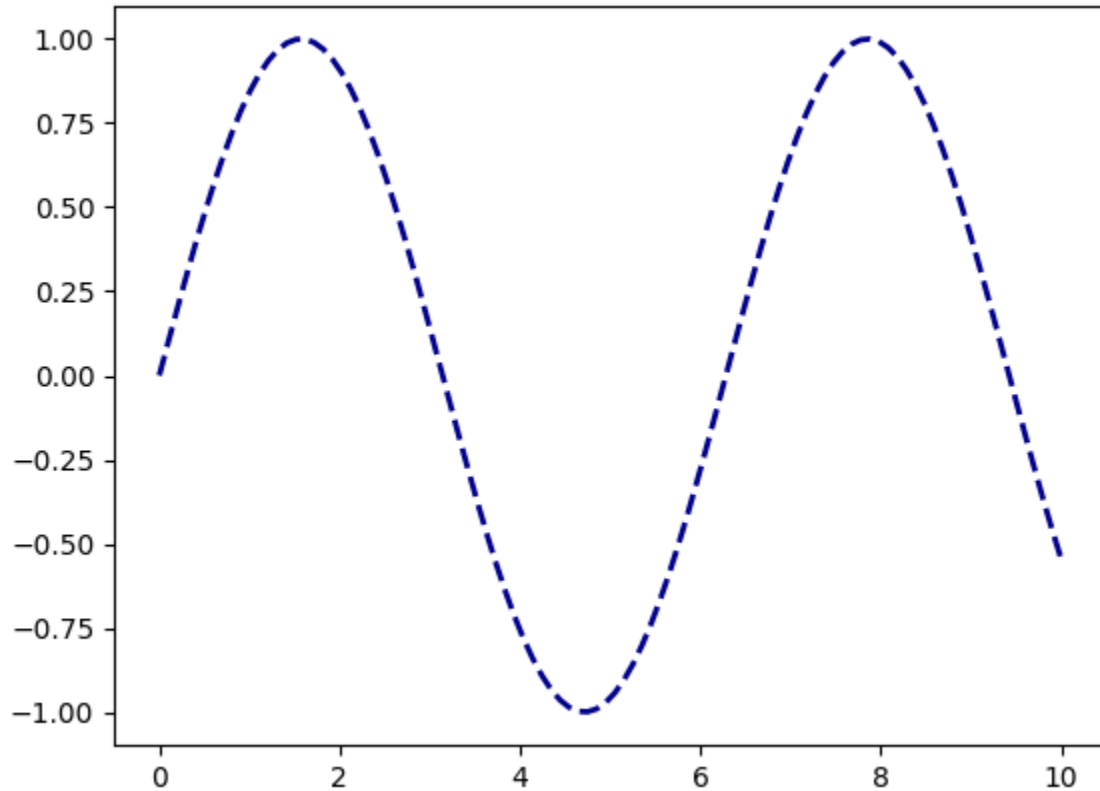
Εδώ η μεταβλητή axes είναι πίνακας που περιέχει δύο αντικείμενα Axes, ένα για κάθε γράφημα. Η αντικειμενοστραφής προσέγγιση μας επιτρέπει τον σαφή διαχωρισμό των στοιχείων κάθε γραφήματος.

7.1.4. Παραμετροποίηση της εμφάνισης

Ένα από τα ισχυρότερα χαρακτηριστικά της Matplotlib είναι η δυνατότητα πλήρους προσαρμογής της εμφάνισης. Συγκεκριμένα, μπορούν να ρυθμιστούν τα χρώματα, οι τύποι και το πάχος των γραμμών, οι γραμματοσειρές, η μορφή των αξόνων αλλά και το ίδιο το στυλ των markers.

Ακολουθεί ένα απλό παράδειγμα με αλλαγή στο στυλ, στο χρώμα, στο πάχος της γραμμής:

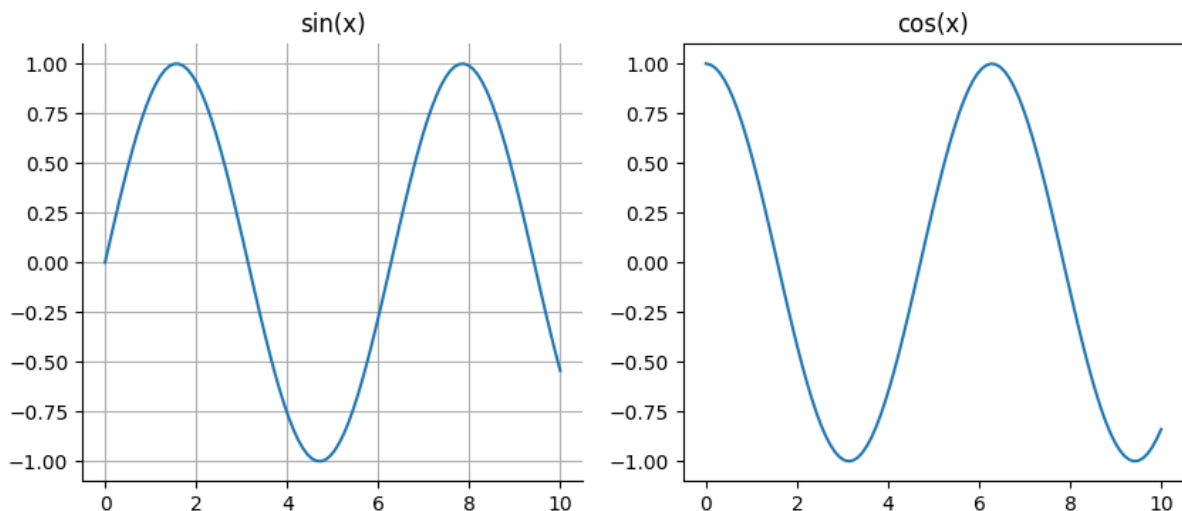
```
plt.plot(x, y, color="darkblue", linestyle="--", linewidth=2);
```



Εικόνα 7.830 Συνάρτηση ημιτόνου με παραμετροποίηση στο στυλ, στο χρώμα, στο πάχος της γραμμής.

Ένα σημαντικό πλεονέκτημα της αντικειμενοστραφούς προσέγγισης είναι ότι μπορούμε να τροποποιούμε άμεσα συγκεκριμένα στοιχεία του κάθε γραφήματος. Ας το δούμε αλλάζοντας μόνο ένα υπογράφημα στο παράδειγμα με τα δύο υπογραφήματα που είδαμε νωρίτερα:

```
x = np.linspace(0, 10, 100)
fig, axes = plt.subplots(1, 2, figsize=(10,4))
axes[0].plot(x, np.sin(x))
axes[0].set_title("sin(x)")
axes[1].plot(x, np.cos(x))
axes[1].set_title("cos(x)")
axes[0].grid(True)
axes[0].spines["top"].set_visible(False)
axes[0].spines["right"].set_visible(False)
plt.show()
```



Εικόνα 7.931 Παραμετροποίηση των υπογραφημάτων της Εικόνας 7.7, ώστε μόνο το αριστερό υπογράφημα να έχει πλέγμα.

7.1.5. Εξαγωγή γραφημάτων

Τα γραφήματα της Matplotlib μπορούν να εξαχθούν σε πολλές μορφές αρχείων:

```
plt.savefig("figure.png", dpi=600)
plt.savefig("figure.pdf", dpi=400)
plt.savefig("figure.svg", dpi=300)
```

Εκτός από τη μορφή, έχουμε έλεγχο και στη ποιότητα, μέσω της παραμέτρου DPI (Dots Per Inch), η οποία μετράει πόσα pixels χωρούν σε μία ίντσα ως ένας δείκτης ανάλυσης (resolution). Χαμηλό DPI (π.χ. 50-70) οδηγεί σε γραμμή που θα φαίνεται "σπασμένη" (pixelated). Από την άλλη πλευρά, υψηλό DPI (π.χ. 300+) οδηγεί σε μία γραμμή που θα είναι απόλυτα λεία και σε καθαρά γράμματα. Φυσικά, όσο μικρότερο το DPI τόσο λιγότερο χώρο αποθήκευσης θα απαιτεί η εικόνα κατά την εξαγωγή της.

7.2. Αισθητική βελτίωση μέσω Seaborn

Η **Seaborn** είναι μία βιβλιοθήκη υψηλότερου επιπέδου που βασίζεται στη Matplotlib. Παρέχει έτοιμες συναρτήσεις για τη δημιουργία σύνθετων γραφημάτων, χωρίς πολλές γραμμές κώδικα και με αισθητικά βελτιωμένες προεπιλογές. Ένα σημαντικό χαρακτηριστικό της Seaborn είναι η άμεση ενσωμάτωσή της με τη βιβλιοθήκη Pandas, γεγονός που διευκολύνει την ανάλυση των δεδομένων.

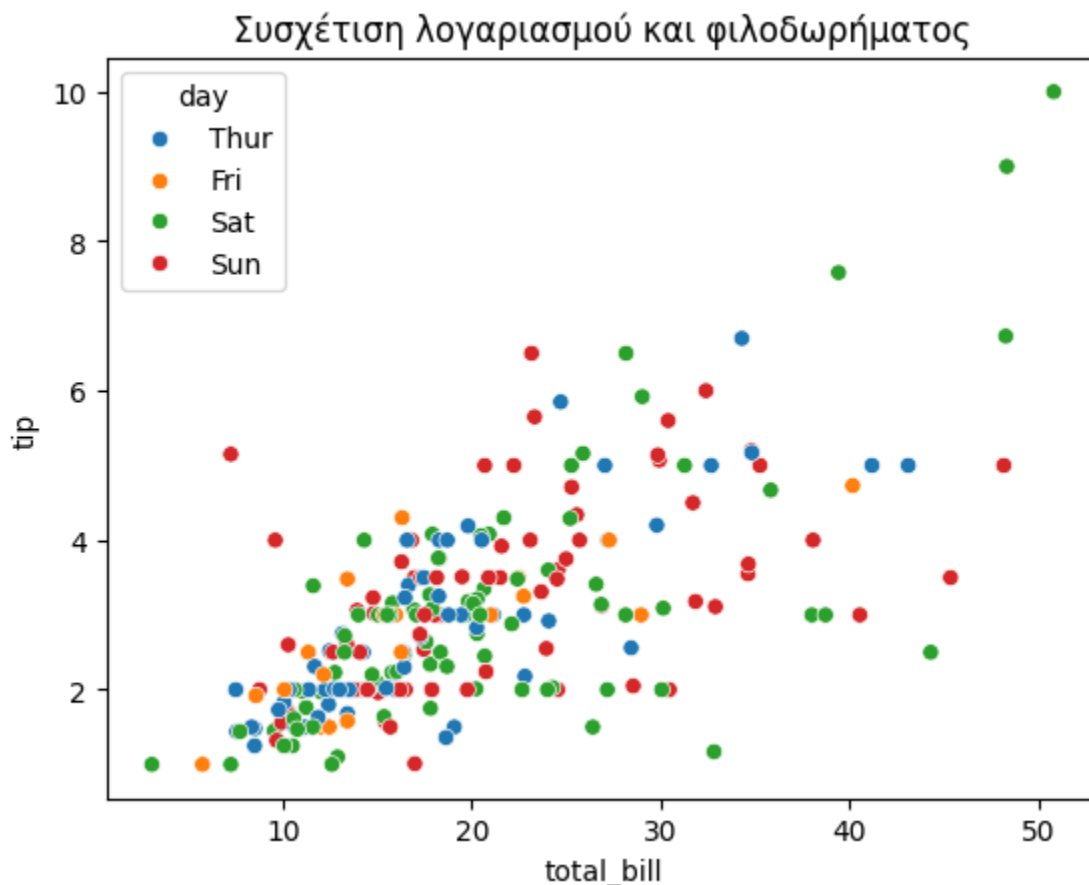
Η Seaborn χρησιμοποιείται ιδιαίτερα συχνά για γραφήματα όπως κατανομές (histograms, KDE plots), boxplots και violin plots, heatmaps και πίνακες συσχέτισης, γραφήματα συσχέτισης μεταβλητών. Στην πράξη, η Seaborn χρησιμοποιείται για γρήγορη εξερευνητική ανάλυση, ενώ η Matplotlib για λεπτομερή προσαρμογή του τελικού γραφήματος.

Ας δούμε ένα παράδειγμα:

```
import seaborn as sns
import matplotlib.pyplot as plt
```

```
tips = sns.load_dataset("tips")

sns.scatterplot(data=tips, x="total_bill", y="tip", hue="day")
plt.title("Συσχέτιση λογαριασμού και φιλοδωρήματος")
plt.show()
```



Εικόνα 7.1032 Συσχέτιση λογαριασμού και φιλοδωρήματος ανά ημέρα μέσω του συνόλου δεδομένων tips.

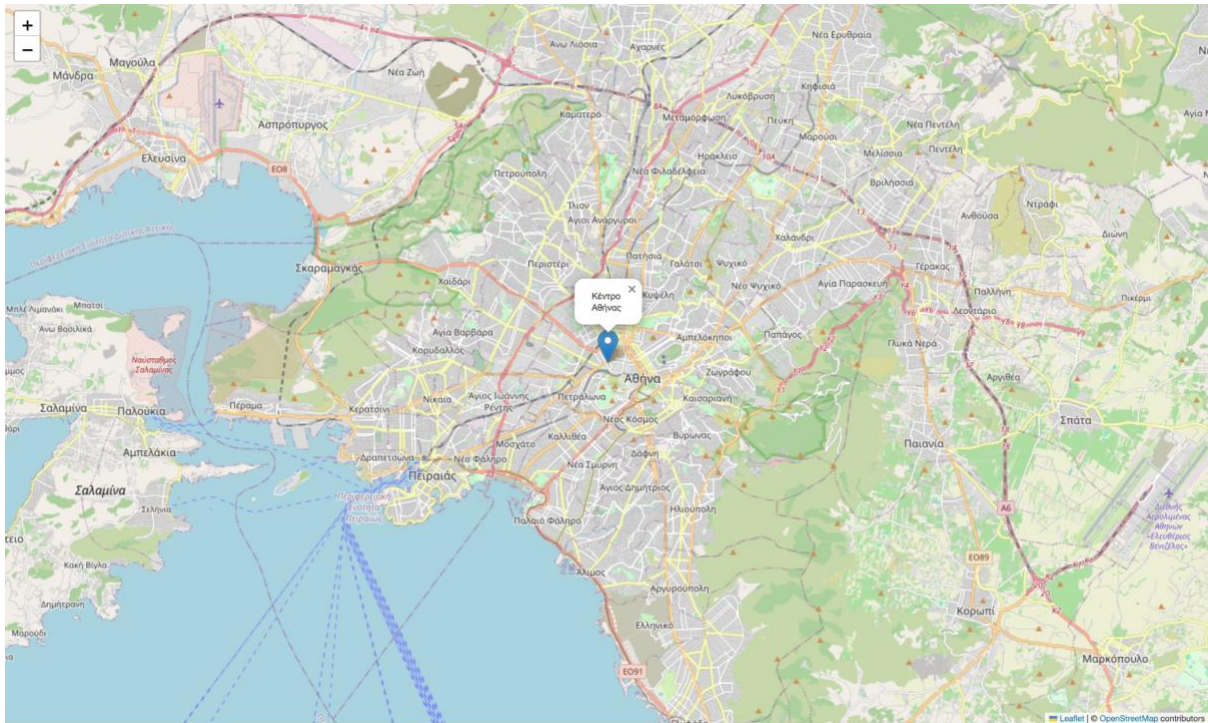
7.3. Χάρτες μέσω Folium

Η βιβλιοθήκη **Folium** βοηθάει τη δημιουργία διαδραστικών γεωχωρικών χαρτών. Βασίζεται στη βιβλιοθήκη (JavaScript) **Leaflet** και επιτρέπει την ενσωμάτωση χαρτών μέσα σε notebooks ή ιστοσελίδες. Γενικότερα, χρησιμοποιείται συχνά όταν τα δεδομένα περιλαμβάνουν γεωγραφικές συντεταγμένες, όπως οι τοποθεσίες συμβάντων, και είναι ιδιαίτερα χρήσιμη σε εφαρμογές γεωχωρικής ανάλυσης και αστικής πληροφορικής. Αυτό συμβαίνει γιατί μπορούμε να δημιουργήσουμε εύκολα διαδραστικούς χάρτες, προσθέτοντας pins ή διαδρομές.

Ένα απλό παράδειγμα ακολουθεί, όπου το αποτέλεσμα είναι ένας χάρτης σε HTML.

```
import folium
map = folium.Map(location=[37.98, 23.72], zoom_start=12)
folium.Marker(
    location=[37.98, 23.72],
    popup="Κέντρο Αθήνας"
).add_to(map)
```

map



Εικόνα 7.1133 Χάρτης με διαδραστικότητα σε HTML μέσω της βιβλιοθήκης Folium.

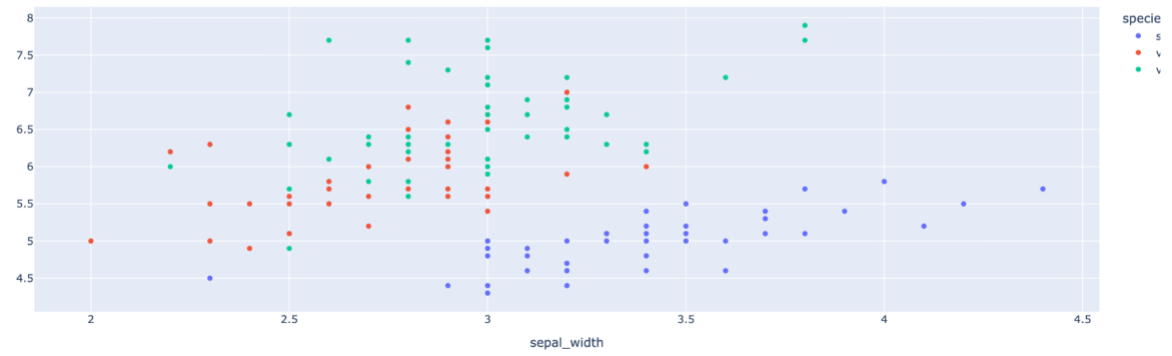
7.4. Διαδραστικότητα μέσω Plotly

Η **Plotly** είναι μία βιβλιοθήκη διαδραστικής οπτικοποίησης, επιτρέποντας τη δημιουργία γραφημάτων τα οποία μπορούν να εξερευνηθούν δυναμικά από τον χρήστη. Τα γραφήματα αυτά υποστηρίζουν λειτουργίες όπως μεγέθυνση, μετακίνηση, εμφάνιση τιμών διαδραστικά. Η Plotly είναι κατάλληλη για διαδραστικά dashboards ή παρουσιάσεις (ειδικά μεγάλων συνόλων) δεδομένων στον ιστό. Χρησιμοποιείται συχνά σε συνδυασμό με το πλαίσιο **Dash**, το οποίο επιτρέπει την ανάπτυξη πλήρων διαδραστικών εφαρμογών δεδομένων.

Ένα απλό παράδειγμα ακολουθεί:

```
import plotly.express as px
df = px.data.iris()
fig = px.scatter(
    df,
    x="sepal_width",
    y="sepal_length",
    color="species",
    title="Διαδραστικό γράφημα Plotly"
)
fig.show()
```

Διαδραστικό γράφημα Plotly



Εικόνα 7.1234 Διαδραστικό γράφημα μέσω της βιβλιοθήκης Plotly για το σύνολο δεδομένων Iris όπου οριζόντια φαίνεται το πλάτος και κάθετα το μήκος του κάθε σέπαλου.