Big Data και Analytics

5η εβδομάδα: Καθαρισμός και προεπεξεργασία δεδομένων

Τμήμα Λογιστικής και Χρηματοοικονομικής

Πλαίσιο της σημερινής συνάντησης

- Διάρκεια μαθήματος: 3 ώρες
- Πρώτο μέρος: γιατί ο καθαρισμός δεδομένων είναι κρίσιμος για λογιστική και χρηματοοικονομική ανάλυση
- Δεύτερο μέρος: βασικά προβλήματα ποιότητας και στάδια προεπεξεργασίας
- Τρίτο μέρος: πρακτικά παραδείγματα με R και συμπληρωματικά με Excel

Στόχος

Να γίνει σαφές ότι η αξιοπιστία κάθε ανάλυσης εξαρτάται άμεσα από την ποιότητα, τη συνέπεια και τη σωστή προετοιμασία των δεδομένων.

Μαθησιακά αποτελέσματα της 5ης εβδομάδας

Με το τέλος της σημερινής διάλεξης οι φοιτητές αναμένεται να μπορούν:

- να αναγνωρίζουν βασικά προβλήματα ποιότητας σε λογιστικά και χρηματοοικονομικά δεδομένα
- να διακρίνουν τον καθαρισμό από τη γενικότερη προεπεξεργασία
- να εφαρμόζουν βασικές τεχνικές χειρισμού ελλειπουσών τιμών, διπλοεγγραφών και ακραίων τιμών
- να κατανοούν γιατί η τεκμηρίωση των μετασχηματισμών είναι απαραίτητη
- να εκτελούν βασικά βήματα καθαρισμού σε R και, όπου χρειάζεται, σε Excel

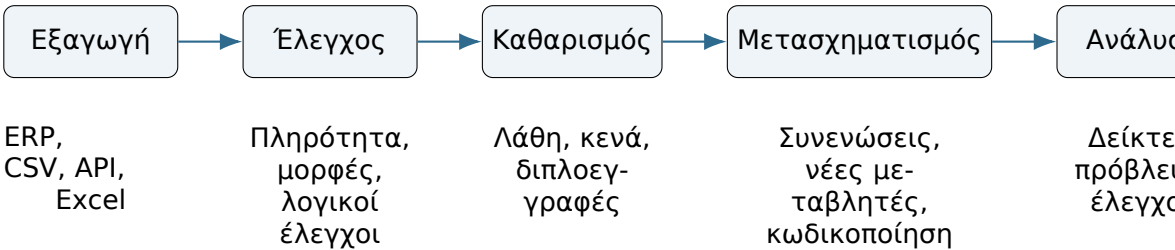
Γιατί το θέμα είναι κομβικό;

- Τα πραγματικά δεδομένα σπάνια είναι έτοιμα για ανάλυση.
- Στη λογιστική και στη χρηματοοικονομική μικρά σφάλματα μπορούν να οδηγήσουν σε σοβαρές στρεβλώσεις.
- Το μεγαλύτερο μέρος ενός έργου analytics δεν είναι το μοντέλο αλλά η προετοιμασία των δεδομένων.
- Η κακή ποιότητα δεδομένων παράγει λανθασμένους δείκτες, εσφαλμένες προβλέψεις και αδύναμα ελεγκτικά ευρήματα.

Κεντρική ιδέα

Garbage in, garbage out.
Αν τα δεδομένα είναι προβληματικά, η ανάλυση θα είναι προβληματική, ανεξάρτητα από το εργαλείο ή τη μέθοδο.

Από τα ακατέργαστα δεδομένα στην ανάλυση



Η προεπεξεργασία δεν είναι βοηθητικό στάδιο. Είναι ο πυρήνας της αναλυτικής αξιοπιστίας.

Τι εννοούμε με τον όρο καθαρισμός δεδομένων;

Ορισμός

Καθαρισμός δεδομένων είναι η διαδικασία εντοπισμού και διόρθωσης λαθών, ελλείψεων, ασυνεπειών και διπλοεγγραφών, ώστε το σύνολο δεδομένων να καταστεί κατάλληλο για ανάλυση.

- Έλεγχος ορθότητας τιμών
- Εναρμόνιση μορφοτύπων
- Απομάκρυνση ή επίλυση ασυνεπειών
- Τεκμηρίωση κάθε παρέμβασης

Τι περιλαμβάνει η προεπεξεργασία δεδομένων;

Καθαρισμός

- Ελλιπείς τιμές
- ασύμβατες μορφές
- διπλοεγγραφές
- ακραίες ή λανθασμένες τιμές

Ευρύτερη προεπεξεργασία

- συνένωση αρχείων
- τυποποίηση κατηγοριών
- δημιουργία νέων μεταβλητών
- αναδιάταξη, φιλτράρισμα και κωδικοποίηση

Η προεπεξεργασία είναι ευρύτερη έννοια από τον απλό καθαρισμό. Προετοιμάζει τα δεδομένα ώστε να απαντούν σε συγκεκριμένο αναλυτικό ερώτημα.

Συνηθισμένα προβλήματα σε λογιστικά και χρηματοοικονομικά δεδομένα

- κενά πεδία σε πελάτες, προμηθευτές, λογαριασμούς ή ημερομηνίες
- διαφορετική εγγραφή του ίδιου στοιχείου, π.χ. ίδιου προμηθευτή
- αρνητικά ποσά εκεί όπου δεν επιτρέπονται από τη λογική της συναλλαγής
- ποσά με λάθος δεκαδικά ή λάθος νόμισμα
- διπλές εγγραφές τιμολογίων ή πληρωμών
- ημερομηνίες σε πολλαπλούς μορφότυπους
- ασυνέπεια μεταξύ υποσυστημάτων, π.χ. πωλήσεων και γενικής λογιστικής

Η λογική του ελέγχου ποιότητας δεδομένων

Διάσταση ποιότητας	Τι ελέγχουμε	Ενδεικτικό παράδειγμα
Πληρότητα	Αν υπάρχουν κενά σε κρίσιμα πεδία	Απουσία ημερομηνίας τιμολογίου
Ακρίβεια	Αν η τιμή αντανακλά την πραγματικότητα	Λάθος ποσό ή νόμισμα
Συνέπεια	Αν το ίδιο στοιχείο εμφανίζεται ομοιόμορφα	Διαφορετική εγγραφή του ίδιου πελάτη
Εγκυρότητα	Αν η τιμή τηρεί κανόνες μορφής ή λογικής	Αρνητική ποσότητα πώλησης χωρίς πίστωση
Μοναδικότητα	Αν υπάρχουν διπλοεγγραφές	Ίδιο τιμολόγιο καταχωρισμένο δύο φορές
Επικαιρότητα	Αν το στοιχείο είναι έγκαιρο και ενημερωμένο	Παλιό υπόλοιπο πελάτη σε τρέχουσα αναφορά

Ελλιπείς τιμές: το πρώτο πρακτικό πρόβλημα

- Μπορεί να οφείλονται σε σφάλμα εισαγωγής, αποτυχία συστήματος ή πραγματική απουσία πληροφορίας.
- Δεν αντιμετωπίζονται όλες με τον ίδιο τρόπο.
- Η επιλογή χειρισμού εξαρτάται από το είδος της μεταβλητής και το αναλυτικό ερώτημα.

Συνήθεις επιλογές

- διαγραφή εγγραφής
- αντικατάσταση με κανόνα
- ειδική κατηγορία unknown
- περαιτέρω διερεύνηση

Προσοχή

Η μηχανική συμπλήρωση κενών τιμών χωρίς τεκμηρίωση μπορεί να αλλοιώσει δείκτες και συμπεράσματα.

Διπλοεγγραφές και οντότητες με πολλαπλή γραφή

- Ίδιος πελάτης ή προμηθευτής μπορεί να εμφανίζεται με διαφορετική γραφή.
- Ίδιο τιμολόγιο μπορεί να έχει καταχωριστεί δύο φορές λόγω διαδικαστικού λάθους.
- Σε ομαδικές αναφορές αυτό οδηγεί σε υπερεκτίμηση πωλήσεων, υποχρεώσεων ή απαιτήσεων.

Ενδεικτικοί έλεγχοι

Αναζήτηση ίδιου αριθμού παραστατικού, ίδιου ποσού, ίδιας ημερομηνίας, παρόμοιας επωνυμίας ή συνδυασμών κλειδιών.

Ακραίες τιμές: σφάλμα ή οικονομικό γεγονός;

- Μια ακραία τιμή δεν είναι πάντοτε λάθος.
- Μπορεί να αποτυπώνει πραγματική συναλλαγή υψηλής αξίας ή ασυνήθιστο γεγονός.
- Η κρίσιμη εργασία είναι η διερεύνηση και όχι η αυτόματη διαγραφή.

Χρήσιμα ερωτήματα

- είναι λογικά εφικτή η τιμή;
- συμφωνεί με το παραστατικό;
- αποτελεί ένδειξη απάτης ή εξαίρεσης;

Τυποποίηση μορφών και κατηγοριών

- Ημερομηνίες: 01/02/2025, 2025-02-01, 1 Feb 2025
- Ποσά: κόμμα ή τελεία ως δεκαδικός διαχωριστής
- Κωδικοί λογαριασμών: διαφορετικό μήκος ή μορφή
- Κατηγορίες: Retail, retail, RET

Σημασία

Χωρίς τυποποίηση, η ομαδοποίηση, η συγχώνευση και η εξαγωγή δεικτών είναι αναξιόπιστες.

Συνένωση δεδομένων από πολλαπλές πηγές

- Συνήθεις συνδυασμοί: γενική λογιστική, πωλήσεις, εισπράξεις, πελάτες, προϋπολογισμοί
- Το κύριο πρόβλημα είναι ο σωστός ορισμός κλειδιών σύνδεσης
- Λανθασμένο join μπορεί να δημιουργήσει τεχνητό πολλαπλασιασμό εγγραφών

Παράδειγμα

Αν ένας πίνακας πελατών περιέχει διπλό κωδικό και συνδεθεί με πίνακα πωλήσεων, οι συνολικές πωλήσεις μπορεί να εμφανιστούν διογκωμένες.

Τεκμηρίωση μετασχηματισμών και αναπαραγωγιμότητα

- Πρέπει να καταγράφεται τι άλλαξε, γιατί άλλαξε και πότε.
- Η τεκμηρίωση είναι κρίσιμη για ελεγκτικά ίχνη και επανάληψη της ανάλυσης.
- Η R προσφέρει ισχυρό πλεονέκτημα επειδή τα βήματα καταγράφονται σε κώδικα.

Καλές πρακτικές

- αρχικό αρχείο μόνο για ανάγνωση
- ξεχωριστό σενάριο καθαρισμού
- σαφείς κανόνες ονοματοδοσίας
- σχόλια στον κώδικα

Γιατί η R είναι το βασικό εργαλείο του μαθήματος;

- Επιτρέπει αναπαραγώγιμη επεξεργασία δεδομένων.
- Υποστηρίζει μεγάλα σύνολα δεδομένων καλύτερα από ένα απλό υπολογιστικό φύλλο.
- Διευκολύνει τον συνδυασμό καθαρισμού, ανάλυσης και οπτικοποίησης στο ίδιο περιβάλλον.
- Ευνοεί συστηματική τεκμηρίωση και έλεγχο των βημάτων.

Ο ρόλος του Excel

Το Excel παραμένει χρήσιμο για γρήγορο έλεγχο, επισκόπηση και απλές διορθώσεις, αλλά δεν είναι το κύριο εργαλείο για σύνθετη και επαναλήψιμη προεπεξεργασία.

Ενδεικτικό παράδειγμα καθαρισμού σε R

```
library(dplyr)
library(readr)
library(stringr)

data <- read_csv("sales_raw.csv")

clean_data <- data %>%
  mutate(customer = str_squish(str_to_upper(customer)),
         invoice_date = as.Date(invoice_date, format = "%d/%m/%Y")) %>%
  filter(!is.na(invoice_id), !is.na(amount)) %>%
  distinct(invoice_id, .keep_all = TRUE)
```

Το παράδειγμα δείχνει τυποποίηση ονομάτων, μετατροπή ημερομηνίας, απομάκρυνση κρίσιμων κενών και εξάλειψη διπλοεγγραφών.

Τι κάνει αυτός ο κώδικας;

Βήμα	Λειτουργία
read_csv()	Εισάγει το αρχείο δεδομένων
str_squish() και str_to_upper()	Τυποποιούν τη γραφή ονομάτων πελατών
as.Date()	Μετατρέπει τις ημερομηνίες σε ενιαίο μορφότυπο
filter()	Αποκλείει εγγραφές χωρίς κρίσιμα πεδία
distinct()	Αφαιρεί διπλά τιμολόγια βάσει invoice_id

Παράδειγμα λογιστικής εφαρμογής

Σενάριο

Μια επιχείρηση θέλει να υπολογίσει το δείκτη μέσου χρόνου είσπραξης ανά πελάτη. Το αρχείο πωλήσεων περιέχει διπλοεγγραφές, κενές ημερομηνίες και διαφορετική γραφή πελατών.

- Αν δεν προηγηθεί καθαρισμός, ο δείκτης θα είναι στρεβλός.
- Ο υπεύθυνος ανάλυσης μπορεί να συμπεράνει λανθασμένα ότι το πρόβλημα αφορά λάθος πελάτες.
- Η διοικητική απόφαση τότε θα στηριχθεί σε εσφαλμένη πληροφόρηση.

Παράδειγμα ελεγκτικής εφαρμογής

- Σε έλεγχο πληρωμών προμηθευτών, οι διπλές εγγραφές είναι πιθανή ένδειξη λάθους ή απάτης.
- Οι ακραίες τιμές μπορεί να υποδηλώνουν έκτακτη συναλλαγή ή παραβίαση ορίων έγκρισης.
- Η προεπεξεργασία είναι προϋπόθεση για αξιόπιστο exception testing.

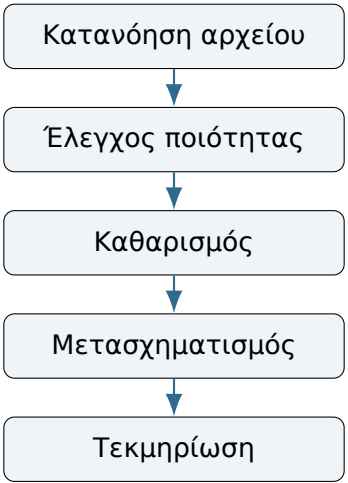
Άμεση σύνδεση

Καθαρά δεδομένα σημαίνουν ισχυρότερα ευρήματα, καλύτερη τεκμηρίωση και πιο πειστικό ελεγκτικό συμπέρασμα.

Συχνά λάθη στην προεπεξεργασία

- Αυτόματη διαγραφή προβληματικών γραμμών χωρίς εξέταση της αιτίας.
- Αλλαγές απευθείας στο αρχικό αρχείο χωρίς ιστορικό ενεργειών.
- Συνένωση πινάκων χωρίς έλεγχο μοναδικότητας κλειδιών.
- Τυποποίηση κατηγοριών χωρίς έλεγχο επιχειρησιακής σημασίας.
- Χρήση του Excel ως μοναδικού εργαλείου σε σύνθετα επαναλαμβανόμενα σενάρια.

Προτεινόμενη διαδικασία εργασίας για τους φοιτητές



Η σωστή σειρά μειώνει λάθη, ενισχύει τη συνέπεια και κάνει την αναλυτική εργασία επαναλήψιμη.

Προτεινόμενη κατανομή του 3ώρου μαθήματος

Μέρος	Περιεχόμενο	Ενδεικτική διάρκεια
1ο μέρος	Θεωρητική εισαγωγή: έννοιες, διαστάσεις ποιότητας, βασικά προβλήματα	60 λεπτά
2ο μέρος	Παραδείγματα από λογιστική, χρηματοοικονομική και έλεγχο	45 λεπτά
3ο μέρος	Εργαστηριακή εφαρμογή σε R, σύντομη αναφορά σε Excel	60 λεπτά
4ο μέρος	Συζήτηση, ανακεφαλαίωση, μικρή άσκηση ή ερωτήσεις	15 λεπτά

Μικρή δραστηριότητα μέσα στην τάξη

Άσκηση

Δίνεται αρχείο με πωλήσεις που περιλαμβάνει κενές ημερομηνίες, διπλά invoice IDs, ασυνεπή ονόματα πελατών και ακραίες τιμές ποσών.

Οι φοιτητές καλούνται να εντοπίσουν:

- ποια προβλήματα αφορούν ποιότητα και ποια προεπεξεργασία
- ποια πεδία είναι κρίσιμα για ανάλυση και δεν επιτρέπουν κενές τιμές
- ποια βήματα καθαρισμού πρέπει να προηγηθούν πριν από οποιονδήποτε υπολογισμό δεικτών

Ανακεφαλαίωση

- Ο καθαρισμός δεδομένων είναι προϋπόθεση και όχι συμπλήρωμα της ανάλυσης.
- Στη λογιστική και στη χρηματοοικονομική η ποιότητα δεδομένων επηρεάζει άμεσα δείκτες, εκτιμήσεις και ελεγκτικά συμπεράσματα.
- Η προεπεξεργασία περιλαμβάνει καθαρισμό, τυποποίηση, συνένωση και μετασχηματισμό.
- Η R είναι το βασικό εργαλείο του μαθήματος επειδή υποστηρίζει αναπαραγώγιμη και τεκμηριωμένη εργασία.

Μετάβαση στην επόμενη εβδομάδα

Η επόμενη ενότητα θα αξιοποιήσει καθαρά δεδομένα για εξερευνητική ανάλυση και οπτικοποίηση.

Ερωτήσεις και συζήτηση

Ευχαριστώ πολύ

Ερωτήσεις, απορίες ή σύντομη συζήτηση πάνω στα παραδείγματα της σημερινής διάλεξης.