Big Data και Analytics

2η εβδομάδα: Πηγές και τύποι δεδομένων στη λογιστική και χρηματοοικονομική

Τμήμα Λογιστικής και Χρηματοοικονομικής

Πλαίσιο της σημερινής συνάντησης

- Πρώτο μέρος: τι είδους δεδομένα χρησιμοποιούν η λογιστική και η χρηματοοικονομική
- Δεύτερο μέρος: από πού προέρχονται τα δεδομένα και πώς αξιολογούνται
- Τρίτο μέρος: δομή δεδομένων, πρακτικά παραδείγματα και πρώτη μικρή επαφή με δεδομένα σε R και Excel

Στόχος

Να κατανοηθεί ότι πριν από κάθε ανάλυση προηγείται η σωστή αναγνώριση των πηγών, του τύπου και των περιορισμών των δεδομένων.

Μαθησιακά αποτελέσματα της 2ης εβδομάδας

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

- να διακρίνουν βασικές κατηγορίες δεδομένων σε λογιστικό και χρηματοοικονομικό περιβάλλον
- να αναγνωρίζουν εσωτερικές και εξωτερικές πηγές δεδομένων
- να εξηγούν τη διαφορά μεταξύ δομημένων, ημιδομημένων και αδόμητων δεδομένων
- να αξιολογούν την καταλληλότητα μιας πηγής δεδομένων για συγκεκριμένο αναλυτικό ερώτημα
- να κατανοούν πώς η R και το Excel διαχειρίζονται διαφορετικούς τύπους δεδομένων

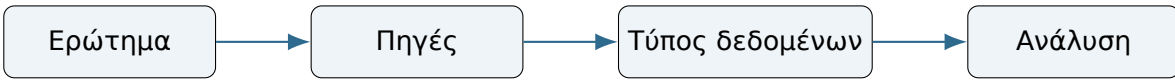
Γιατί ξεκινάμε από τις πηγές δεδομένων;

- Ένα επιχειρησιακό πρόβλημα μπορεί να απαιτεί συνδυασμό πολλών πηγών δεδομένων.
- Η ποιότητα της ανάλυσης περιορίζεται από την ποιότητα και την καταλληλότητα των εισροών/δεδομένων.
- Στη λογιστική και χρηματοοικονομική πολλά κρίσιμα δεδομένα είναι διάσπαρτα σε διαφορετικά συστήματα.
- Η ερμηνεία εξαρτάται από το πλαίσιο παραγωγής των δεδομένων.

Κεντρική ιδέα

Δεν υπάρχει καλή αναλυτική χωρίς σωστή χαρτογράφηση των δεδομένων που διαθέτουμε και αυτών που λείπουν.

Από το επιχειρησιακό ερώτημα στην πηγή δεδομένων



Τι θέλουμε να απαντήσουμε;

Από πού θα αντλήσουμε στοιχεία;

Πίνακες, κείμενα, ροές, αρχεία

Περιγραφή, διάγνωση, πρόβλεψη

Το σωστό ταίριασμα ερωτήματος, πηγής και μορφής δεδομένων είναι απαραίτητη προϋπόθεση για ένα αξιόπιστο αποτέλεσμα.

Κύριες διακρίσεις δεδομένων

Διάκριση	Κατηγορία Α	Κατηγορία Β
Προέλευση	Εσωτερικά δεδομένα από συστήματα της επιχείρησης	Εξωτερικά δεδομένα από αγορές, τράπεζες, δημόσιες πηγές
Μορφή	Δομημένα δεδομένα σε πίνακες και πεδία	Ημιδομημένα ή αδόμητα δεδομένα σε κείμενα, αρχεία
Χρονική διάσταση	Ιστορικά δεδομένα	Δεδομένα σχεδόν πραγματικού χρόνου
Χρήση	Λειτουργικά ή λογιστικά δεδομένα	Στρατηγικά, αγοραία ή συμπεριφορικά δεδομένα

Εσωτερικές πηγές δεδομένων

Λογιστική και λειτουργία

- γενική λογιστική
- αναλυτική λογιστική
- αγορές και προμήθειες
- αποθήκη και αποθέματα
- μισθοδοσία
- προϋπολογισμοί

Πληροφοριακά συστήματα

- ERP
- CRM
- treasury systems
- πληροφοριακά συστήματα ελέγχου
- αρχεία καταγραφής πρόσβασης και εγκρίσεων
- dashboards και εσωτερικές αναφορές

Τι Είναι ένα Dashboard;

Τα dashboards (πίνακες ελέγχου) είναι οπτικά εργαλεία που συγκεντρώνουν, οργανώνουν και παρουσιάζουν σημαντικά δεδομένα και μετρικές (KPIs) σε μία ενιαία οθόνη, ώστε να επιτρέπουν τη γρήγορη παρακολούθηση και λήψη αποφάσεων.

Κύρια Χαρακτηριστικά

1. **Οπτικοποίηση**
   Χρήση γραφημάτων, πινάκων, δεικτών και χρωμάτων για εύκολη κατανόηση
2. **Πραγματικός χρόνος**
   Ενημέρωση δεδομένων σε πραγματικό χρόνο ή σχεδόν πραγματικό χρόνο
3. **Προσαρμογή**
   Δυνατότητα προσαρμογής ανάλογα με τις ανάγκες του χρήστη
4. **Διαδραστικότητα**
   Φίλτρα, drill-down και δυνατότητα εμβάθυνσης στα δεδομένα

Τι Είναι ένα Dashboard;

Τύποι Dashboards

1. **Επιχειρηματικά (Business):** Παρακολούθηση πωλήσεων, εσόδων, απόδοσης τμημάτων
2. **Αναλυτικά (Analytics):** Βαθιά ανάλυση δεδομένων, τάσεις, προβλέψεις
3. **Λειτουργικά (Operational):** Παρακολούθηση διαδικασιών σε πραγματικό χρόνο
4. **Στρατηγικά (Strategic):** Μακροπρόθεσμοι στόχοι και KPIs για τη διοίκηση

Εξωτερικές πηγές δεδομένων

- τραπεζικά δεδομένα και κινήσεις λογαριασμών
- επιτόκια, συναλλαγματικές ισοτιμίες και τιμές αγοράς
- οικονομικές καταστάσεις τρίτων, reports και ανακοινώσεις
- εκθέσεις αναλυτών και δεδομένα πιστοληπτικής αξιολόγησης
- ειδησεογραφικές ροές και κανονιστικές ανακοινώσεις
- δεδομένα κοινωνικών δικτύων και εναλλακτικά δεδομένα αγοράς

Παρατήρηση

Η χρηματοοικονομική ανάλυση σπάνια στηρίζεται μόνο σε εσωτερικά δεδομένα. Συνήθως απαιτεί συνδυασμό λογιστικών, αγοραίων και μακροοικονομικών πληροφοριών.

Δομημένα, ημιδομημένα και αδόμητα δεδομένα

Δομημένα

- πίνακες
- στήλες
- αριθμητικά πεδία
- κωδικοί λογαριασμών

Ημιδομημένα

- csv
- json
- xml
- αρχεία με μεταδεδομένα

Αδόμητα

- emails
- συμβάσεις
- αναφορές ελέγχου
- ειδησεογραφικά κείμενα

Στην πράξη, τα πιο απαιτητικά προβλήματα προκύπτουν όταν χρειάζεται να συνδυαστούν και οι τρεις μορφές.

Παράδειγμα αντιστοίχισης τύπου δεδομένων και χρήσης

Τύπος δεδομένων	Ενδεικτικά παραδείγματα	Τυπική χρήση
Αριθμητικά	ποσά, υπόλοιπα, επιτόκια, αποδόσεις	υπολογισμοί, δείκτες, προβλέψεις
Κατηγορικά	κωδικός πελάτη, χώρα, κλάδος, τύπος δαπάνης	ομαδοποίηση, ταξινόμηση, τμηματοποίηση
Χρονικά	ημερομηνίες τιμολόγησης, πληρωμής, λήξης	τάσεις, εποχικότητα, καθυστερήσεις
Κειμένου	σχόλια ελεγκτή, περιγραφές συναλλαγών, ειδήσεις	ποιοτική ερμηνεία, εξόρυξη κειμένου

Κλασικά σύνολα δεδομένων στη λογιστική

- εγγραφές ημερολογίου
- ισοζύγια και λογαριασμοί καθολικού
- τιμολόγια αγορών και πωλήσεων
- απαιτήσεις, υποχρεώσεις και ληξιπρόθεσμα υπόλοιπα
- δεδομένα κόστους και κέντρων ευθύνης
- αρχεία απογραφής και αποθεμάτων

Αναλυτική αξία

Τα δεδομένα αυτά υποστηρίζουν συμφωνίες, ελέγχους εξαιρέσεων, κοστολόγηση, ανάλυση αποκλίσεων και έγκαιρη προειδοποίηση.

Κλασικά σύνολα δεδομένων στη χρηματοοικονομική

- ταμειακές ροές
- χρηματοδοτήσεις και αποπληρωμές
- στοιχεία χαρτοφυλακίου
- αποδόσεις και διακυμάνσεις
- στοιχεία πιστωτικού κινδύνου
- αγοραίες τιμές
- καμπύλες επιτοκίων
- συναλλαγματικές ισοτιμίες
- δεδομένα μακροοικονομικού περιβάλλοντος
- εξωτερικές αξιολογήσεις και ratings

Η χρηματοοικονομική ανάλυση απαιτεί συστηματικά δεδομένα πολλαπλών χρονικών κλιμάκων και διαφορετικής προέλευσης.

Παράδειγμα 1: ανάλυση καθυστερημένων απαιτήσεων

Ερώτημα

Ποιοι πελάτες εμφανίζουν αυξημένο κίνδυνο καθυστέρησης ή αθέτησης;

- Εσωτερικά δεδομένα: τιμολόγια, ημερομηνίες λήξης, ιστορικό πληρωμών, πιστωτικά όρια.
- Εξωτερικά δεδομένα: κλάδος, οικονομική κατάσταση πελάτη, ειδήσεις ή rating.
- Τύποι δεδομένων: αριθμητικά, κατηγορικά, χρονικά και πιθανώς κειμένου.

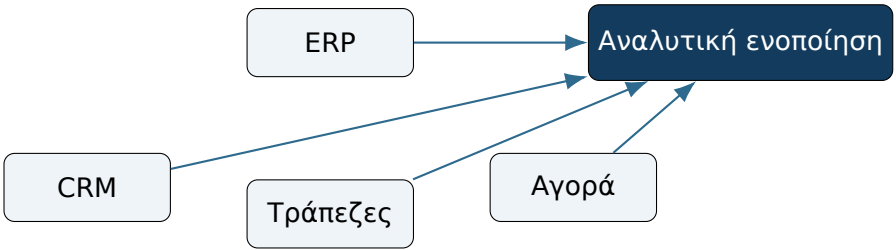
Παράδειγμα 2: ανάλυση κερδοφορίας προϊόντων

Ερώτημα

Γιατί υποχώρησε το περιθώριο κέρδους σε συγκεκριμένη κατηγορία προϊόντων;

- Δεδομένα πωλήσεων, εκπτώσεων και επιστροφών.
- Δεδομένα κόστους πρώτων υλών, μεταφοράς και παραγωγής.
- Δεδομένα αποθεμάτων και προμηθευτών.
- Πιθανή σύνδεση με εξωτερικά δεδομένα αγοράς ή συναλλαγματικών μεταβολών.

Πρόβλημα της διάσπασης των δεδομένων



Το πρακτικό πρόβλημα δεν είναι μόνο η ύπαρξη δεδομένων, αλλά η σύνδεση διαφορετικών συστημάτων με συνέπεια και σωστή αντιστοίχιση.

Πότε μια πηγή δεδομένων είναι κατάλληλη;

- Όταν απαντά στο σωστό επιχειρησιακό ερώτημα.
- Όταν το επίπεδο ανάλυσης είναι επαρκώς λεπτομερές.
- Όταν η χρονική κάλυψη είναι σωστή.
- Όταν γνωρίζουμε ποιος τη δημιούργησε και για ποιο σκοπό.
- Όταν μπορούμε να ελέγξουμε ποιότητα, πληρότητα και συνέπεια.

Σημείο προσοχής

Δεν είναι κάθε διαθέσιμο dataset κατάλληλο για κάθε χρήση. Η προσβασιμότητα δεν ισοδυναμεί με αναλυτική καταλληλότητα.

Συνήθη προβλήματα στις πηγές δεδομένων

- ελλιπείς τιμές
- διαφορετικοί κωδικοί για το ίδιο αντικείμενο
- ασυνέπειες μορφοποίησης
- πολλαπλές εκδόσεις του ίδιου αρχείου
- λανθασμένες ημερομηνίες
- έλλειψη μοναδικού αναγνωριστικού
- μεταβολές ορισμών στο χρόνο
- περιορισμένη τεκμηρίωση προέλευσης

Αυτά τα προβλήματα εξηγούν γιατί η προετοιμασία των δεδομένων θα αποτελέσει ξεχωριστή ενότητα του μαθήματος.

Δεδομένα κατά επίπεδο ανάλυσης

Επίπεδο	Παράδειγμα	Αναλυτική χρησιμότητα
Συναλλαγή	μία πληρωμή ή ένα τιμολόγιο	έλεγχος εξαιρέσεων, ανίχνευση απάτης
Πελάτης ή προμηθευτής	ιστορικό σχέσης, όρια, συμπεριφορά	τμηματοποίηση, πιστωτική πολιτική
Τμήμα ή μονάδα	πωλήσεις, κόστος, προϋπολογισμός	αξιολόγηση απόδοσης, αποκλίσεις
Επιχείρηση ή όμιλος	συνολικά οικονομικά μεγέθη	στρατηγική, χρηματοδοτική και επενδυτική ανάλυση

Δεδομένα σε Excel και R

Εργαλείο	Τι εξυπηρετεί καλύτερα	Περιορισμός
Excel	πρώτη επισκόπηση, φίλτρα, pivot tables, βασικές συμφωνίες και αναφορές	γίνεται δύσχρηστο όσο αυξάνονται όγκος, επαναληψιμότητα και πολυπλοκότητα
R	εισαγωγή, συνένωση, καθαρισμός, αναπαραγώγιμη ανάλυση και οπτικοποίηση	απαιτεί εξοικείωση με σύνταξη και οργανωμένη ροή εργασίας

Μικρό παράδειγμα φόρτωσης δεδομένων σε R

```r
# Δεδομένα πελατών από csv
customers <- read.csv("customers.csv")

# Δεδομένα πληρωμών από Excel
# install.packages("readxl")
library(readxl)
payments <- read_excel("payments.xlsx")

# Πρώτος έλεγχος δομής
str(customers)
str(payments)

# Βασική σύνοψη
summary(customers)
```

Τι μας ενδιαφέρει εδώ;

Όχι ο προγραμματισμός ως αυτοσκοπός, αλλά η ικανότητα να αναγνωρίζουμε τι περιέχει κάθε αρχείο και πώς θα το χρησιμοποιήσουμε αναλυτικά.

Πρακτική δραστηριότητα

Σενάριο

Μια επιχείρηση θέλει να εξετάσει αν η αύξηση καθυστερημένων απαιτήσεων σχετίζεται με αλλαγές στον τύπο πελατών και στον κλάδο δραστηριότητας.

- Ποια εσωτερικά δεδομένα θα ζητούσατε;
- Ποια εξωτερικά δεδομένα θα προσθέτατε;
- Ποια πεδία είναι αριθμητικά, ποια κατηγορικά και ποια χρονικά;
- Τι θα βλέπατε πρώτα σε Excel και τι θα οργανώνατε σε R;

1. Εσωτερικά Δεδομένα

Ανίχνευση στοιχείων από το ERP και το CRM της επιχείρησης

Αρχεία Λογιστηρίου & Πωλήσεων

- **Master Data:** ΑΦΜ, Κλάδος (ΚΑΔ), Γεωγραφική Ζώνη, Ημ/νία Έναρξης Συνεργασίας.
- **Συναλλακτικά Δεδομένα:** Αρ. Τιμολογίου, Ημ/νία Έκδοσης, *Due Date*, Ημ/νία Πληρωμής.
- **Πιστωτική Πολιτική:** Ιστορικό ορίων πίστωσης και εσωτερικό scoring.

Ποιοτικά Δεδομένα

- Σημειώσεις εισπράξεων.
- Ιστορικό επικοινωνίας και αμφισβητήσεων.

Target Variable (Y): Υπολογισμός των *Ημερών Καθυστέρησης* (Actual Date - Due Date).

2. Εξωτερικά Δεδομένα

Εμπλουτισμός για τη διάκριση Ιδιοσυγκρατικού vs Συστημικού Κινδύνου

Μακροοικονομικό Περιβάλλον

- **Κόστος Χρήματος:** Επιτόκια δανεισμού και πληθωρισμός.
- **Κλαδικοί Δείκτες:** Δείκτες κύκλου εργασιών ανά ΚΑΔ (ΕΛΣΤΑΤ).
- **Οικονομικό Κλίμα:** Δείκτες εμπιστοσύνης (ΙΟΒΕ).

Δεδομένα Αγοράς

- **Φερεγγυότητα:** Δεδομένα πτωχεύσεων και δικαστικών πράξεων (ΓΕΜΗ).
- **Market Benchmarks:** Μέσοι όροι ημερών πληρωμής ανά κλάδο.

”Η καθυστέρηση οφείλεται στον πελάτη ή σε κρίση του κλάδου του;”

3. Τυπολογία Δεδομένων (Data Dictionary)

Οργάνωση πεδίων για τη στατιστική ανάλυση

Πεδίο	Τύπος	Χρήση στην Ανάλυση
Ποσό Τιμολογίου	Αριθμητικό	Μέγεθος οικονομικής έκθεσης.
Ημέρες Καθυστέρησης	Αριθμητικό	**Μεταβλητή Στόχος (*Y*)**.
Τύπος Πελάτη	Κατηγορικό	Σύγκριση ομάδων (π.χ. B2B vs Gov).
Κλάδος (ΚΑΔ)	Κατηγορικό	Βασικός ερμηνευτικός παράγοντας.
Πιστωτική Βαθμίδα	Διατακτικό	Ιεραρχική αξιολόγηση κινδύνου.

4. Επιλογή Εργαλείου: Excel vs R

MS Excel: Διερεύνηση

- **Pivot Tables:** Γρήγορη σύνοψη καθυστερήσεων ανά κλάδο.
- **Data Viz:** Χρονοσειρές για τον εντοπισμό τάσεων.
- **Reporting:** Dashboards άμεσης πληροφόρησης.

R / Python: Τεκμηρίωση

- **ANOVA / `lm()`**: Στατιστική σημαντικότητα διαφορών.
- **Time-Series**: Απομόνωση εποχικότητας από την τάση.
- **Inference**: Πρόβλεψη πιθανότητας αθέτησης πληρωμής.

Συμπέρασμα: Το **Excel** απαντά στο *«Τι συμβαίνει;»*, η **R** στο *«Γιατί συμβαίνει και τι θα γίνει μετά;»*

Τρία βασικά σημεία που πρέπει να μείνουν

1. Το ίδιο αναλυτικό πρόβλημα μπορεί να απαιτεί πολλά και διαφορετικά δεδομένα.
2. Η προέλευση, η δομή και η χρονική διάσταση των δεδομένων επηρεάζουν άμεσα την ανάλυση.
3. Η επιλογή κατάλληλης πηγής προηγείται του μοντέλου και του εργαλείου.

Συνέχεια του μαθήματος

Στην επόμενη εβδομάδα θα περάσουμε από την αναγνώριση πηγών στη διακυβέρνηση και στην ποιότητα των δεδομένων, δηλαδή στο πώς εξασφαλίζεται ότι τα δεδομένα είναι αξιόπιστα και ελέγξιμα.

Προετοιμασία για την επόμενη εβδομάδα

- Να εντοπιστούν δύο παραδείγματα εσωτερικών και δύο παραδείγματα εξωτερικών πηγών δεδομένων από πραγματικό επιχειρησιακό περιβάλλον.
- Να γίνει μια πρώτη καταγραφή πιθανών προβλημάτων ποιότητας σε αρχεία csv ή Excel.
- Να υπάρχει έτοιμο το περιβάλλον R/RStudio για χρήση στις επόμενες ασκήσεις.

Επόμενο μάθημα

Διακυβέρνηση δεδομένων, ποιότητα δεδομένων και κανονιστική συμμόρφωση.

Ενδεικτική βιβλιογραφία για αφετηρία

- Provost, F. and Fawcett, T., Data Science for Business.
- Shmueli, G. et al., Data Mining for Business Analytics.
- Appelbaum, D., Kogan, A. and Vasarhelyi, M., αρθρογραφία για accounting data analytics και audit data.

Ευχαριστώ