Εργασία εξαμήνου: Instacart

Αυτή η σελίδα είναι ο **οδηγός** σας για την εργασία: ξεκινά από το μηδέν (τι είναι το Kaggle και το notebook), σας οδηγεί βήμα-βήμα στα πρώτα υποχρεωτικά ερωτήματα με **σημεία ελέγχου** για να ξέρετε ότι είστε σωστά, και μετά σας αφήνει χώρο να πειραματιστείτε πάνω στο θέμα που θα διαλέξετε. Ό,τι χρειάζεστε το έχετε ήδη μάθει: φίλτρα, groupby, merge και γραφήματα.

ΜΕ ΜΙΑ ΜΑΤΙΑ

Ατομικά ή σε ομάδες 2 ατόμων. 50% του βαθμού. Παραδοτέα: notebook + παρουσίαση 10′. Πλατφόρμα: Kaggle Notebooks, ξεκινάτε με Copy & Edit από το starter notebook: **kaggle.com/code/evangeloskalampokis/starter-notebook**. Δεδομένα: Instacart 2017 (5 αρχεία CSV, ήδη συνδεδεμένα στο notebook). Εργαστήρια βοήθειας: Πέμ 10 Δεκ 2026 (Μέρος Α) και Πέμ 17 Δεκ 2026 (Μέρος Β). Παρουσιάσεις και παράδοση: **Παρ 8 Ιαν 2027 και Παρ 15 Ιαν 2027**. Δήλωση ομάδας και θέματος στη φόρμα μέχρι Πέμ 3 Δεκ 2026.

1. Πριν ξεκινήσεις: Kaggle και notebooks

Μέχρι τώρα τρέχατε κώδικα μέσα σε αυτές τις σελίδες. Για την εργασία θα δουλέψετε σε ένα **notebook** *(Jupyter notebook)*: ένα έγγραφο που αποτελείται από **κελιά** *(cells)*. Κάθε κελί είναι είτε **κώδικας** που τρέχει και δείχνει το αποτέλεσμά του ακριβώς από κάτω, είτε **κείμενο** *(Markdown)* όπου εξηγείτε τι κάνετε και τι βρήκατε. Κώδικας, αποτελέσματα, γραφήματα και σχόλια μαζί, με τη σειρά: γι' αυτό το notebook είναι ταυτόχρονα το εργαλείο σας και η αναφορά σας.

Το **Kaggle** *(kaggle.com)* είναι μια δωρεάν πλατφόρμα για δεδομένα και notebooks που τρέχει στον browser, σε υπολογιστή της Google: δεν εγκαθιστάτε τίποτα, το pandas και το Matplotlib είναι ήδη εκεί, τα δεδομένα της εργασίας είναι ήδη «κουμπωμένα» στο notebook, και κάθε notebook έχει στη διάθεσή του 30 GB μνήμης, αρκετή για τα 32 εκατομμύρια γραμμές του Instacart. Η δουλειά σας σώζεται στον λογαριασμό σας στο Kaggle.

ΤΟ NOTEBOOK ΤΗΣ ΕΡΓΑΣΙΑΣ

Ξεκινάτε από εδώ: kaggle.com/code/evangeloskalampokis/starter-notebook. Ανοίξτε το, πατήστε **Copy & Edit** και έχετε δικό σας αντίγραφο με τα δεδομένα ήδη συνδεδεμένα.

Πώς ανοίγω το notebook της εργασίας

1. Πήγαινε στο kaggle.com → **Register** και φτιάξε λογαριασμό (με Google ή με email, δύο λεπτά). **Αν είστε δύο, φτιάχνετε και οι δύο λογαριασμό**, αλλά δουλεύετε σε ένα notebook, όχι σε δύο αντίγραφα.
2. Άνοιξε το **starter notebook** του μαθήματος (kaggle.com/code/evangeloskalampokis/starter-notebook) και πάτησε **Copy & Edit** (πάνω δεξιά). Ανοίγει ένα δικό σου αντίγραφο στον editor, με τα δεδομένα του Instacart ήδη συνδεδεμένα: δες το panel **Input** δεξιά.

3. Πάνω αριστερά, κάνε κλικ στο όνομα του notebook και μετονόμασέ το, π.χ. `BAP Instacart Ομάδα 07`. Σώζεται αυτόματα ως πρόχειρο *(draft)* στον λογαριασμό σου, στη σελίδα **Your Work**.
4. Κάνε κλικ στο πρώτο κελί κώδικα και πάτησε **Shift + Enter**. Την πρώτη φορά περιμένει λίγα δευτερόλεπτα για να «ξυπνήσει» τον υπολογιστή *(session)* και μετά τυπώνει το αποτέλεσμα από κάτω.
5. Πάτησε **+ Code** για νέο κελί κώδικα ή **+ Markdown** για κελί κειμένου (τα κουμπιά εμφανίζονται όταν πας το ποντίκι ανάμεσα σε δύο κελιά). Στο κελί κειμένου γράφεις με Markdown: `# Τίτλος`, `## Υπότιτλος`, `**έντονα**`, `- λίστα`.
6. **Save Version** (πάνω δεξιά) → **Quick Save**: κρατά μια «φωτογραφία» της δουλειάς σου. Κάντε το κάθε φορά που τελειώνετε κάτι. Ο αριθμός δίπλα στο κουμπί δείχνει τις εκδόσεις, και από εκεί γυρίζετε σε παλιότερη αν χρειαστεί.
7. **Share** (πάνω δεξιά) → αν είστε δύο, πρόσθεσε τον συνεργάτη σου με το Kaggle username του ως *collaborator* με δικαίωμα **Can edit**. Το Kaggle ζητά επαλήθευση με κινητό για αυτό (Settings → Phone verification, ένα SMS). Δουλεύετε **με τη σειρά**, όχι ταυτόχρονα: ο ένας κάνει Save Version πριν συνεχίσει ο άλλος.

2. Η εργασία με μια ματιά

Η **Instacart** είναι πλατφόρμα online παραγγελιών από σούπερ μάρκετ στις ΗΠΑ, με παράδοση στο σπίτι *(online grocery delivery)*. Το 2017 δημοσίευσε ανωνυμοποιημένα δεδομένα: **3,4 εκατομμύρια παραγγελίες** από 206 χιλιάδες πελάτες, με τα προϊόντα κάθε καλαθιού, την ημέρα και ώρα της παραγγελίας, και αν κάθε προϊόν ήταν επαναγορά. Είναι από τα πιο γνωστά σύνολα δεδομένων στην επιχειρηματική αναλυτική, και ακριβώς το είδος δεδομένων που έχει κάθε λιανέμπορος.

Ο ρόλος σας: είστε η ομάδα αναλυτικής της Instacart. Η διοίκηση θέλει να καταλάβει *πώς παραγγέλνουν οι πελάτες* και τι μπορεί να κάνει με αυτή τη γνώση. Κάθε ομάδα απαντά πρώτα σε ένα κοινό σύνολο ερωτημάτων (βήμα 4), διαλέγει μετά ένα **θέμα εμβάθυνσης** (βήμα 5) και καταλήγει σε **ευρήματα και μία πρόταση** προς τη διοίκηση.

ΟΜΑΔΕΣ

Ατομικά ή σε **ομάδες 2 ατόμων**. Ομάδα και θέμα δηλώνονται σε μια φόρμα μέχρι την ημερομηνία του χρονοδιαγράμματος. Όποιος δεν δηλώσει ομάδα, εργάζεται ατομικά. Στη συνέχεια λέμε «ομάδα» και για όσους δουλεύουν μόνοι τους.

ΒΑΘΜΟΣ

Η εργασία είναι το **50%** του τελικού βαθμού. Το υπόλοιπο 50% είναι η γραπτή εξέταση. Στις ομάδες των 2, και τα δύο μέλη παίρνουν τον ίδιο βαθμό, εκτός αν η ομάδα ζητήσει διαφορετικά με αιτιολόγηση.

ΠΑΡΑΔΟΤΕΑ

Ένα notebook (.ipynb) που τρέχει από την αρχή ως το τέλος χωρίς σφάλμα, και μία παρουσίαση 10 λεπτών (5–6 διαφάνειες). Λεπτομέρειες στο βήμα 6.

ΕΡΓΑΛΕΙΑ

Kaggle Notebooks, pandas, Matplotlib. Ό,τι έχουμε κάνει στο μάθημα. Τίποτα επιπλέον δεν απαιτείται (αλλά επιτρέπεται).

Τα τρία μέρη της εργασίας

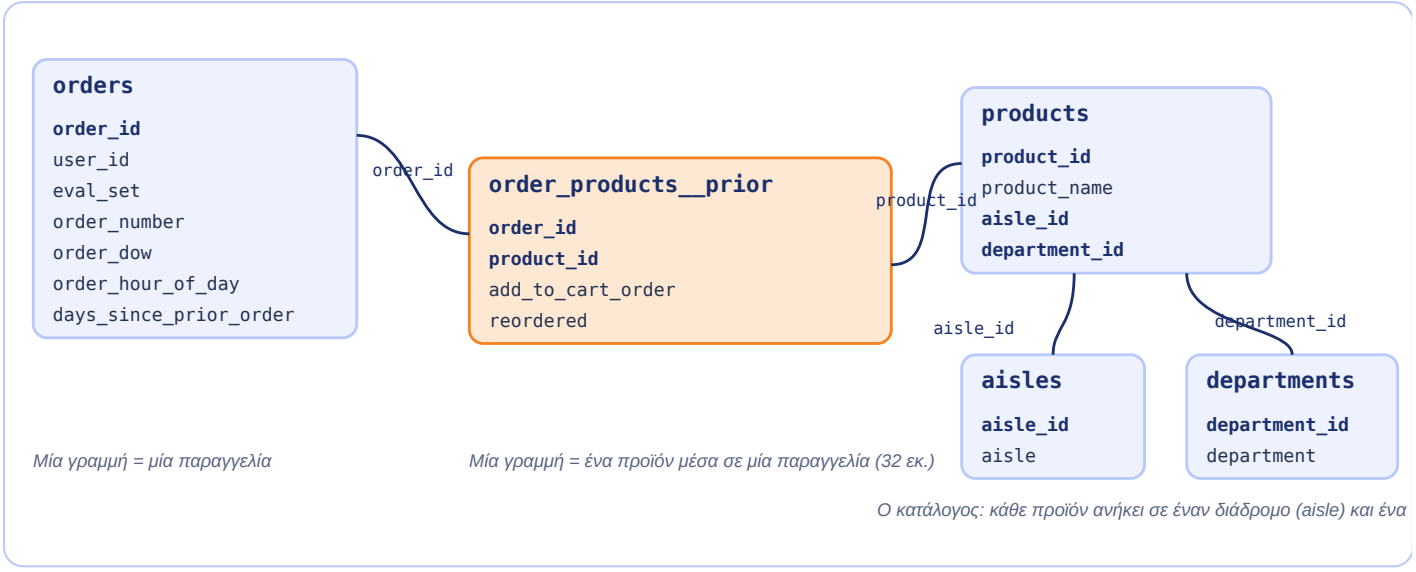
Μέρος	Τι κάνετε	Πού βρίσκεται η καθοδήγηση	Βάρος
`Α. Εκκίνηση (κοινό για όλους)`	Φόρτωση, έλεγχος, merge και 6 υποχρεωτικά ερωτήματα με σημεία ελέγχου	Βήμα 4	30%
`Β. Εμβάθυνση (ανά ομάδα)`	Ένα θέμα από το μενού, 3 καθοδηγούμενα ερωτήματα + τουλάχιστον 1 δικό σας	Βήμα 5	45%
`Γ. Επικοινωνία`	Καθαρό notebook με σχόλια, γραφήματα, συμπεράσματα, παρουσίαση 10'	Βήμα 6	25%

ΤΙ ΔΕΝ ΕΙΝΑΙ Η ΕΡΓΑΣΙΑ

Δεν είναι διαγωνισμός machine learning, ούτε «όσο περισσότερος κώδικας τόσο καλύτερα». Είναι: σωστά νούμερα, καθαρά γραφήματα, και μια απάντηση σε επιχειρηματική ερώτηση που θα καταλάβαινε ένας διευθυντής χωρίς να διαβάσει κώδικα.

3. Τα δεδομένα

Πέντε αρχεία CSV που συνδέονται μεταξύ τους με «κλειδιά» *(ids)*, όπως σε μια βάση δεδομένων. Η κεντρική ιδέα: οι **παραγγελίες** είναι ένας πίνακας, τα **προϊόντα κάθε παραγγελίας** άλλος, και τα **προϊόντα** με τις κατηγορίες τους ένας τρίτος. Για να απαντήσετε «πόσες μπανάνες πουλήθηκαν την Κυριακή» χρειάζεστε και τους τρεις, ενωμένους με `merge`.



Τα κλειδιά (με έντονα) είναι οι στήλες που ενώνουν τους πίνακες με `merge`. Ο κεντρικός πίνακας είναι το order_products: από εκεί ξεκινούν τα περισσότερα ερωτήματα.

Αρχείο	Γραμμές	Στήλες	Τι είναι
`orders.csv`	3.421.083	order_id, user_id, eval_set, order_number, order_dow, order_hour_of_day, days_since_prior_order	Μία γραμμή ανά παραγγελία: ποιος, πότε (ημέρα 0–6, ώρα 0–23), πόσες μέρες από την προηγούμενη (NaN στην πρώτη παραγγελία κάθε πελάτη)
`order_products__prior.csv`	32.434.489	order_id, product_id, add_to_cart_order, reordered	Μία γραμμή ανά προϊόν μέσα σε παραγγελία: η σειρά που μπήκε στο καλάθι και αν ήταν επαναγορά (1) ή πρώτη φορά (0)
`products.csv`	49.688	product_id, product_name, aisle_id, department_id	Ο κατάλογος προϊόντων
`aisles.csv`	134	aisle_id, aisle	Ο «διάδρομος» (υποκατηγορία), π.χ. fresh fruits, yogurt
`departments.csv`	21	department_id, department	Το τμήμα, π.χ. produce, dairy eggs, snacks

ΤΙ ΑΓΝΟΟΥΜΕ

Στο dataset υπάρχει και ένα αρχείο `order_products__train.csv`, και στο `orders.csv` μια στήλη `eval_set`. Αφορούν τον διαγωνισμό πρόβλεψης του 2017. **Δεν τα χρησιμοποιούμε.** Τα προϊόντα των παραγγελιών είναι στο `order_products__prior.csv`, και μόνο εκεί κοιτάμε.

Πού είναι τα αρχεία

Σε ένα **Kaggle Dataset** (Instacart Market Basket Analysis, kaggle.com/datasets/psparks/instacart-market-basket-analysis), ήδη συνδεδεμένο με το starter notebook: το βλέπετε στο panel **Input** δεξιά. Δεν χρειάζεται upload, mount ή άδεια. Το πρώτο κελί κώδικα του starter notebook τυπώνει τις διαδρομές των αρχείων και το δεύτερο τα φορτώνει με το `read_csv` που ξέρετε, σε πέντε DataFrames: `orders`, `prior`, `products`, `aisles`, `departments`. Τρέξτε τα δύο κελιά και ελέγξτε ότι το τελευταίο `print` δίνει `(3421083, 7) (32434489, 4) (49688, 4) (134, 2) (21, 2)`. Αν ναι, είστε έτοιμοι για το Μέρος Α.

ΣΥΜΒΟΥΛΕΣ ΓΙΑ ΤΑ ΜΕΓΑΛΑ ΑΡΧΕΙΑ

(1) Γράψε και δοκίμασε κάθε ανάλυση πρώτα σε ένα δείγμα, π.χ. `mini = prior.head(1_000_000)` (τρέχει σε δευτερόλεπτα). Όταν δουλεύει, άλλαξε τη μεταβλητή σε `prior`. (2) Κάνε merge **μόνο τις στήλες που χρειάζεσαι**: `products[["product_id", "product_name", "department_id"]]`. (3) Το Kaggle δίνει 30 GB μνήμης, που φτάνουν για όλα τα αρχεία μαζί. Αν παρ' όλα αυτά η συνεδρία «πέσει» (ο δείκτης RAM δεξιά γεμίσει), ξεκίνα πάλι (Run → Restart & clear cell outputs) και φόρτωσε λιγότερες στήλες. (4) Μη φορτώνεις το prior δύο φορές. (5) Αν αφήσεις το notebook ανοιχτό χωρίς να το ακουμπάς για αρκετή ώρα, η συνεδρία κλείνει: ο κώδικας και οι εκδόσεις μένουν, οι μεταβλητές χάνονται. Ξανατρέχεις τα κελιά από την αρχή.

4. Μέρος Α: βήματα εκκίνησης με σημεία ελέγχου

Έξι ερωτήματα, κοινά για όλες τις ομάδες. Για το καθένα δίνουμε τι να κάνετε, ποια εργαλεία του μαθήματος χρειάζονται, και ένα **σημείο ελέγχου** *(checkpoint)*: ένα νούμερο που πρέπει να βγάλετε κι εσείς. Αν βγάλετε άλλο, κάτι πήγε στραβά (συνήθως λάθος αρχείο, λάθος merge ή ξεχασμένο φίλτρο) και το βρίσκετε πριν χτίσετε πάνω του. Τσεκάρετε κάθε βήμα όταν το ολοκληρώσετε. Η λίστα θυμάται την πρόοδό σας σε αυτόν τον browser.

☐ A1. Φόρτωσε και επιθεώρησε τα πέντε αρχεία

`read_csv, shape, head, info`

Για κάθε πίνακα: `shape`, `head()`, `info()`. Σε ένα κελί κειμένου γράψτε με δικά σας λόγια τι είναι μία γραμμή κάθε πίνακα (η «μονάδα»). Αυτή η πρόταση θα σας σώσει από τα μισά λάθη αργότερα.

Σημείο ελέγχου
`orders.shape → (3421083, 7)`, `prior.shape → (32434489, 4)`, `products.shape → (49688, 4)`, aisles 134 γραμμές, departments 21. Πόσοι πελάτες; `orders["user_id"].nunique() → 206209`.

☐ A2. Πότε παραγγέλνουν οι πελάτες;

`value_counts, groupby, bar chart`

Παραγγελίες ανά ημέρα της εβδομάδας (`order_dow`, 0–6) και ανά ώρα (`order_hour_of_day`). Ένα ραβδόγραμμα για το καθένα. Ποια είναι η ώρα αιχμής; Ποιες δύο μέρες ξεχωρίζουν; (Η Instacart δεν λέει ποια μέρα είναι το 0. Τι υποθέτετε, και γιατί;)

Σημείο ελέγχου
Η ώρα με τις περισσότερες παραγγελίες είναι η **10** (λίγο πάνω από τις 11, 14 και 15). Οι μέρες 0 και 1 έχουν η καθεμία πάνω από 580.000 παραγγελίες, οι υπόλοιπες κάτω από 470.000.

☐ A3. Πόσο συχνά παραγγέλνουν;

`groupby, size, describe, histogram`

Παραγγελίες ανά πελάτη: `orders.groupby("user_id").size()` και `describe()`. Ιστόγραμμα. Μετά η κατανομή του `days_since_prior_order`: ποιες δύο τιμές ξεχωρίζουν και τι σημαίνουν για τη συνήθεια των πελατών; (Προσοχή: η τιμή 30 είναι «30 ή περισσότερες». Τα δεδομένα είναι κομμένα εκεί.)

Σημείο ελέγχου
Κάθε πελάτης έχει από **4** έως **100** παραγγελίες, μέσος όρος **≈ 16,6**. Στο days_since_prior_order οι δύο κορυφές είναι στο **7** και στο **30**.

☐ A4. Ένωσε τους πίνακες σε έναν

`merge, shape πριν/μετά`

Φτιάξτε έναν πίνακα `items` με μία γραμμή ανά προϊόν-σε-παραγγελία και δίπλα το όνομα, τον διάδρομο και το τμήμα: `prior → products → aisles → departments`, με `merge(..., on=...)`. **Ελέγξτε τον αριθμό γραμμών πριν και μετά από κάθε merge**: πρέπει να μένει ίδιος. Αν αυξηθεί, υπάρχουν διπλά κλειδιά. Αν μειωθεί, χάνονται γραμμές (inner join).

Σημείο ελέγχου
`items.shape[0] → 32434489` , ίδιο με το prior, και μετά τα τρία merge. Οι στήλες: order_id, product_id, add_to_cart_order, reordered, product_name, aisle_id, department_id, aisle, department.

☐ A5. Τα δημοφιλέστερα προϊόντα και τμήματα

`value_counts, groupby, sort_values, head, barh`

Τα 15 προϊόντα με τις περισσότερες πωλήσεις (γραμμές στο items) και το μερίδιο κάθε τμήματος (`department`) στο σύνολο των προϊόντων που πουλήθηκαν. Οριζόντιο ραβδόγραμμα για τα προϊόντα, ραβδόγραμμα ή πίτα για τα τμήματα. Τι παρατηρείτε για το είδος των προϊόντων στην κορυφή;

Σημείο ελέγχου
Πρώτο προϊόν: **Banana** (472.565 φορές), δεύτερο **Bag of Organic Bananas** (379.450), τρίτο **Organic Strawberries**. Το τμήμα **produce** έχει το μεγαλύτερο μερίδιο, περίπου **29%** των προϊόντων, με δεύτερο το dairy eggs.

☐ A6. Καλάθι και επαναγορές

`groupby, size, mean σε 0/1`

Μέγεθος καλαθιού: προϊόντα ανά παραγγελία (`items.groupby("order_id").size()`), μέσος και διάμεσος, ιστόγραμμα. Ποσοστό επαναγορών: ο μέσος όρος της στήλης `reordered` (είναι 0/1, άρα ο μέσος είναι ποσοστό, όπως μάθαμε). Μετά το ίδιο ποσοστό **ανά τμήμα**: ποια τμήματα ξαναγοράζονται περισσότερο και ποια λιγότερο, και γιατί βγάζει νόημα;

Σημείο ελέγχου
Μέσο καλάθι **≈ 10,1** προϊόντα, διάμεσος **8**. Συνολικό ποσοστό επαναγορών **≈ 0,59**. Ανά τμήμα, το dairy eggs και το produce είναι ψηλά (πάνω από 0,6), ενώ τμήματα όπως personal care και pantry χαμηλά (κάτω από 0,4).

ΟΤΑΝ ΤΕΛΕΙΩΣΕΤΕ ΤΟ ΜΕΡΟΣ Α

Έχετε ήδη έναν ενωμένο πίνακα, έξι γραφήματα και μια πρώτη εικόνα του πελάτη της Instacart. Γράψτε σε ένα κελί κειμένου **τρεις προτάσεις** με τα πιο ενδιαφέροντα που είδατε. Από αυτές τις προτάσεις συνήθως βγαίνει και το θέμα εμβάθυνσης.

5. Μέρος Β: διαλέξτε θέμα και πειραματιστείτε

Έξι κατευθύνσεις. Κάθε μία έχει μια **επιχειρηματική ερώτηση**, τρία **καθοδηγούμενα ερωτήματα** που πρέπει να απαντήσετε (με τα εργαλεία που χρειάζονται), και ιδέες για να το πάτε **παραπέρα**. Υποχρεωτικό: τα τρία καθοδηγούμενα συν **τουλάχιστον ένα δικό σας ερώτημα**, που θα διατυπώσετε και θα αιτιολογήσετε εσείς. Εκεί είναι ο χώρος σας για πειραματισμό. Εκεί διαφέρουν οι πολύ καλές εργασίες. Κάντε κλικ σε ένα θέμα για να το ανοίξετε. Όταν αποφασίσετε, δηλώστε το (μαζί με την ομάδα σας) στη φόρμα δήλωσης μέχρι την προθεσμία του χρονοδιαγράμματος.

`Θέμα 1` Πότε παραγγέλνουν οι πελάτες, και τι σημαίνει για τη λειτουργία

«Πότε χρειαζόμαστε περισσότερους διανομείς και πότε αξίζει να στέλνουμε προσφορές;»

1. Συγκεντρωτικός πίνακας παραγγελιών **ημέρα × ώρα** (`pivot_table` με aggfunc="count") και απεικόνισή του ως heatmap (`plt.imshow` ή `pcolormesh`). Πού είναι η «καυτή ζώνη»;
2. Διαφέρει το **μέγεθος καλαθιού** ανά ώρα ή ανά ημέρα; (merge items με orders, groupby, mean)
3. Διαφέρουν τα **τμήματα** που αγοράζονται το πρωί από το βράδυ; (ποσοστό κάθε τμήματος ανά ζώνη ώρας: πρωί / μεσημέρι / απόγευμα / βράδυ, με `pd.cut` ή φίλτρα)

Παραπέρα: παραγγέλνουν οι «πιστοί» πελάτες (πολλές παραγγελίες) σε άλλες ώρες από τους νέους; Οι επαναγορές γίνονται πιο πολύ Σαββατοκύριακο; Προτείνετε ένα πρόγραμμα βαρδιών με βάση τον πίνακα ημέρα × ώρα.

Θέμα 2 Το καλάθι: μέγεθος, σύνθεση, και τι μπαίνει πρώτο

«Τι χαρακτηρίζει ένα μεγάλο καλάθι, και τι βάζουν πρώτο οι πελάτες;»

1. Κατανομή του μεγέθους καλαθιού και σύγκριση **μικρών** (≤ 5) με **μεγάλων** (≥ 15) καλαθιών: ποια τμήματα κυριαρχούν σε κάθε ομάδα;
2. Τα προϊόντα που μπαίνουν **πρώτα** στο καλάθι (`add_to_cart_order == 1`): top 10, και πώς διαφέρουν από τα top 10 συνολικά;
3. Σχέση **θέσης στο καλάθι** και **επαναγοράς**: είναι τα πρώτα προϊόντα πιο συχνά επαναγορές; (groupby add_to_cart_order, mean reordered, γραμμή)

Παραπέρα: από πόσα τμήματα αγοράζει ένα τυπικό καλάθι (`nunique` ανά παραγγελία); Μεγαλώνει το καλάθι όσο ο πελάτης ωριμάζει (ανά `order_number`); Ποια προϊόντα εμφανίζονται σχεδόν πάντα σε μεγάλα καλάθια;

Θέμα 3 Προϊόντα που ξαναγοράζονται, και προϊόντα «μιας φοράς»

«Ποια προϊόντα κρατούν τους πελάτες και ποια δοκιμάζονται και εγκαταλείπονται;»

1. Ποσοστό επαναγοράς **ανά προϊόν** (mean του reordered) **μαζί** με το πλήθος αγορών (`agg` με δύο στατιστικά). Ταξινομήστε, αλλά **μόνο για προϊόντα με τουλάχιστον 100 αγορές**: γιατί είναι απαραίτητο αυτό το φίλτρο;
2. Σχέση δημοτικότητας και επαναγοράς: διάγραμμα διασποράς (scatter) πλήθος αγορών × ποσοστό επαναγοράς. Ξαναγοράζονται περισσότερο τα δημοφιλή;
3. Προϊόντα **«μιας φοράς»**: με χαμηλό ποσοστό επαναγοράς παρότι αγοράζονται πολύ. Ποια είναι και σε ποια τμήματα ανήκουν;

Παραπέρα: ποσοστό επαναγοράς ανά διάδρομο (aisle) μέσα σε ένα τμήμα. Πόσα προϊόντα κάθε πελάτη είναι «δικά του» (τα ξαναγοράζει σε πάνω από 50% των παραγγελιών του); Μια πρόταση για το «Buy it again» της εφαρμογής.

Θέμα 4 Τμήματα και διάδρομοι: ανατομία ενός τμήματος

«Πώς συγκρίνονται τα τμήματα, και τι συμβαίνει μέσα σε ένα από αυτά;»

1. Προφίλ ανά τμήμα με `agg`: πλήθος προϊόντων στον κατάλογο, πλήθος πωλήσεων, ποσοστό επαναγοράς, μέση θέση στο καλάθι. Ταξινομημένος πίνακας-σύνοψη.
2. Διαλέξτε **ένα τμήμα** (π.χ. produce, snacks, beverages) και κατεβείτε στους **διαδρόμους** του: ποιοι πουλάνε, ποιοι ξαναγοράζονται, ποιοι έχουν πολλά προϊόντα αλλά λίγες πωλήσεις;
3. Το τμήμα σας στον χρόνο: ανά ημέρα και ώρα, διαφέρει από το σύνολο;

Παραπέρα: ποια τμήματα αγοράζονται μαζί συχνά (ποσοστό καλαθιών που περιέχουν και A και B); Ο «κανόνας 80/20»: τι ποσοστό προϊόντων κάνει το 80% των πωλήσεων σε κάθε τμήμα (`cumsum`); Πρόταση για τον κατάλογο: ποιοι διάδρομοι θέλουν περισσότερα ή λιγότερα προϊόντα;

Θέμα 5 Πελάτες: συχνότητα, πιστότητα, τμηματοποίηση

«Ποιοι είναι οι καλύτεροι πελάτες μας, και πώς συμπεριφέρονται διαφορετικά;»

1. Προφίλ ανά πελάτη με `agg` πάνω στο orders: πλήθος παραγγελιών, μέσες μέρες ανάμεσα σε παραγγελίες, συνηθισμένη ώρα. Κατανομές (ιστογράμματα).
2. **Τμήματα πελατών** με `pd.cut` πάνω στο πλήθος παραγγελιών (π.χ. 4–9 νέοι, 10–29 τακτικοί, 30+ πιστοί). Πόσοι σε κάθε τμήμα, και τι ποσοστό των παραγγελιών φέρνει το καθένα;
3. Διαφέρει το **καλάθι** ανάμεσα στα τμήματα πελατών (μέγεθος, ποσοστό επαναγοράς, κορυφαία τμήματα προϊόντων); Χρειάζεται merge items → orders.

Παραπέρα: ο ρυθμός παραγγελιών (μέσες μέρες) αλλάζει καθώς ο πελάτης «ωριμάζει»; Πελάτες «Σαββατοκύριακου» vs «καθημερινής». Μια πρόταση για πρόγραμμα επιβράβευσης με βάση τα τμήματα.

`θέμα 6` Organic ή όχι: ένα ερώτημα κατηγορίας

«Πόσο μεγάλη είναι η αγορά των βιολογικών μέσα στην Instacart, και ποιος την αγοράζει;»

1. Ορίστε στήλη `is_organic` από το όνομα προϊόντος (`products["product_name"].str.contains("Organic")`). Τι ποσοστό των προϊόντων του καταλόγου και τι ποσοστό των **πωλήσεων** είναι organic; Γιατί διαφέρουν τα δύο ποσοστά;
2. Ποσοστό organic **ανά τμήμα** (mean της is_organic ανά department, στο items). Πού είναι ισχυρά τα βιολογικά και πού απουσιάζουν;
3. Ξαναγοράζονται τα organic περισσότερο ή λιγότερο από τα συμβατικά; (groupby is_organic, mean reordered, μετά το ίδιο μέσα σε κάθε τμήμα)

Παραπέρα: «organic πελάτες»: ποσοστό organic ανά πελάτη, κατανομή, και πώς παραγγέλνουν διαφορετικά. Ζεύγη προϊόντων organic / συμβατικού (π.χ. Banana vs Bag of Organic Bananas): ποιο προτιμάται πότε; Άλλες λέξεις-κλειδιά: "Gluten Free", "Low Fat".

Το σχέδιο της ομάδας

Συμπληρώστε το και πατήστε «Φτιάξε το πρώτο κελί»: παράγει το κείμενο Markdown για το **πρώτο κελί** του notebook σας (ομάδα, θέμα, ερώτημα, πλάνο). Αντιγράψτε το σε ένα κελί κειμένου στο Kaggle. Ένα notebook που ξεκινά με το ερώτημά του είναι ήδη μισοτελειωμένο.

6. Μέρος Γ: το notebook, η παρουσίαση, τα κριτήρια

Η δομή του notebook

Ένα κελί κειμένου με τίτλο για κάθε ενότητα, και κάτω από κάθε αποτέλεσμα **μία πρόταση** που λέει τι δείχνει. Ο αναγνώστης (και ο βαθμολογητής) πρέπει να καταλαβαίνει το notebook διαβάζοντας μόνο τα κελιά κειμένου και τα γραφήματα.

#	Ενότητα	Περιεχόμενο
0	Τίτλος, ομάδα, ερώτημα	Το κελί που φτιάξατε στο βήμα 5
1	Φόρτωση δεδομένων	read_csv, shape κάθε πίνακα, η «μονάδα» κάθε πίνακα σε μία πρόταση
2	Μέρος Α: εκκίνηση	Τα έξι ερωτήματα, το καθένα με κώδικα, αποτέλεσμα, γράφημα όπου ζητείται, μία πρόταση συμπέρασμα
3	Μέρος Β: εμβάθυνση	Τα τρία καθοδηγούμενα ερωτήματα και το δικό σας. Κάθε γράφημα με τίτλο, άξονες και μία πρόταση «τι βλέπουμε»

4	Ευρήματα και πρόταση	3–5 κουκκίδες με τα ευρήματα (με νούμερα) και μία πρόταση προς τη διοίκηση που προκύπτει από αυτά
5	Περιορισμοί	Τι δεν ξέρουμε από αυτά τα δεδομένα (π.χ. τιμές, ποια μέρα είναι το 0, μόνο ΗΠΑ 2017) και τι θα θέλατε να είχατε

Η παρουσίαση: 10 λεπτά, 5–6 διαφάνειες

1. **Ερώτημα** (1 διαφάνεια): ποια επιχειρηματική ερώτηση απαντάτε και γιατί έχει σημασία. Όχι «τι είναι η Instacart» για 3 λεπτά.
2. **Δεδομένα** (1): τι χρησιμοποιήσατε, σε δύο προτάσεις, και ένα νούμερο που εντυπωσιάζει.
3. **Τρία ευρήματα** (3): ένα γράφημα ανά διαφάνεια, με τίτλο που *λέει το εύρημα* («Οι Κυριακές 10–11 είναι η αιχμή», όχι «Παραγγελίες ανά ώρα»).
4. **Πρόταση και περιορισμοί** (1): τι θα κάνατε αν ήσασταν η διοίκηση, και τι δεν ξέρετε.

Στις ομάδες των 2 μιλούν και οι δύο. Κώδικας στην παρουσίαση: **καθόλου**, εκτός αν μία γραμμή είναι το ίδιο το εύρημα.

Κριτήρια βαθμολόγησης

Κριτήριο	Βάρος	Τι σημαίνει «άριστα»
`Ορθότητα Μέρους Α`	30%	Και τα έξι ερωτήματα σωστά (τα σημεία ελέγχου βγαίνουν), merge ελεγμένα, κώδικας που τρέχει από την αρχή χωρίς σφάλμα
`Εμβάθυνση Μέρους Β`	30%	Τα τρία ερωτήματα απαντημένα σωστά και ερμηνευμένα. Σωστή χρήση φίλτρων ελάχιστου πλήθους όπου χρειάζεται
`Το δικό σας ερώτημα`	15%	Καλά διατυπωμένο, αιτιολογημένο, απαντημένο με τα κατάλληλα εργαλεία. Δείχνει σκέψη, όχι μόνο κώδικα
`Γραφήματα και κείμενο στο notebook`	15%	Κάθε γράφημα με τίτλο και άξονες, κάθε αποτέλεσμα με μία πρόταση. Ο αναγνώστης καταλαβαίνει χωρίς να διαβάσει κώδικα
`Παρουσίαση`	10%	Στον χρόνο, ένα εύρημα ανά διαφάνεια, καθαρή πρόταση, όλοι μιλούν, απαντήσεις στις ερωτήσεις

Λίστα ελέγχου πριν την παράδοση

- [ ] Save Version → Save & Run All (Commit): τρέχει από την αρχή ως το τέλος χωρίς σφάλμα

- [ ] Τα σημεία ελέγχου του Μέρους Α βγαίνουν
- [ ] Κάθε merge έχει έλεγχο γραμμών πριν/μετά
- [ ] Κάθε γράφημα έχει τίτλο και ονόματα αξόνων
- [ ] Κάτω από κάθε αποτέλεσμα υπάρχει μία πρόταση κειμένου
- [ ] Υπάρχει ενότητα «Ευρήματα και πρόταση» με νούμερα
- [ ] Το δικό μας ερώτημα είναι διατυπωμένο και αιτιολογημένο
- [ ] Το notebook έχει το όνομα ή τα ονόματά σας στο πρώτο κελί
- [ ] Κατεβάσαμε το .ipynb (File → Download Notebook) και το ανεβάσαμε στο eClass
- [ ] Το notebook έγινε Public και ο σύνδεσμός του είναι στο σχόλιο της υποβολής

ΣΥΧΝΑ ΛΑΘΗ ΠΟΥ ΚΟΣΤΙΖΟΥΝ

Ποσοστά επαναγοράς σε προϊόντα με 3 αγορές (100% από 3 δεν λέει τίποτα). Merge που διπλασιάζει γραμμές χωρίς να το προσέξει κανείς. Γράφημα χωρίς τίτλο που «μιλάει μόνο του». Συμπεράσματα που δεν στηρίζονται σε νούμερο. Notebook που τρέχει μόνο με τη σειρά που το έφτιαξε η ομάδα.

7. Χρονοδιάγραμμα και παράδοση

Πότε	Τι
`Πέμ 26 Νοε 2026`	Ανάθεση της εργασίας (εβδ. 9). Διαβάστε αυτή τη σελίδα, φτιάξτε λογαριασμό στο Kaggle, ανοίξτε το starter notebook με Copy & Edit, φορτώστε τα δεδομένα.
`Πέμ 3 Δεκ 2026`	Προθεσμία δήλωσης ομάδας και θέματος εμβάθυνσης στη φόρμα.
`Πέμ 10 Δεκ 2026`	Εργαστήριο στην τάξη: **Μέρος Α** με βοήθεια (εβδ. 11). Έρχεστε με το notebook ανοιχτό και τα δεδομένα φορτωμένα.
`Πέμ 17 Δεκ 2026`	Εργαστήριο στην τάξη: **Μέρος Β** με βοήθεια (εβδ. 12). Έρχεστε με το Μέρος Α ολοκληρωμένο και τα σημεία ελέγχου σωστά.
`Παρ 8 Ιαν 2027`	Παρουσιάσεις, πρώτο μέρος. **Παράδοση notebook και παρουσίασης στο eClass την ίδια μέρα, πριν το μάθημα**, για τις ομάδες που παρουσιάζουν.
`Παρ 15 Ιαν 2027`	Παρουσιάσεις, δεύτερο μέρος. Παράδοση την ίδια μέρα, πριν το μάθημα, για τις ομάδες που παρουσιάζουν. Η σειρά των ομάδων ανακοινώνεται στο eClass.

Πώς παραδίδετε

1. **Δήλωση ομάδας και θέματος**: Δήλωση ομάδας και θέματος, forms.gle/KMfZtqSLonQhkwgK7. Μία απάντηση ανά ομάδα, από το ένα μέλος, με τα δύο ονόματα και το θέμα εμβάθυνσης. Όποιος εργάζεται ατομικά δηλώνει

μόνο το θέμα.

2. **Notebook**: στο αντίγραφο του starter notebook που δουλεύετε στο Kaggle, πρώτα **Save Version → Save & Run All (Commit)**: τρέχει όλο το notebook από την αρχή σε καθαρή συνεδρία, ώστε να είστε σίγουροι ότι δουλεύει. Μετά **File → Download Notebook**: κατεβαίνει το `.ipynb`. Όνομα αρχείου `BAP_Instacart_ΟμάδαXX.ipynb`. Ανεβάστε το στο eClass (Εργασίες → Εργασία εξαμήνου), **μία υποβολή ανά ομάδα**. Την ίδια μέρα κάντε το notebook **Public** (Share → Public, και Save αν το ζητήσει) και γράψτε τον σύνδεσμό του στο σχόλιο της υποβολής: έτσι βλέπουμε και τα αποτελέσματα όπως έτρεξαν στο Kaggle. **Όχι νωρίτερα**: μέχρι την παράδοση το notebook μένει Private, ορατό μόνο στην ομάδα σας.
3. **Πότε**: notebook και παρουσίαση (PDF ή PowerPoint) ανεβαίνουν στο eClass **την ημέρα της παρουσίασής σας, πριν το μάθημα**. Η σειρά των ομάδων (8 ή 15 Ιανουαρίου) ανακοινώνεται στο eClass.

Βοήθεια

Στις εβδομάδες 11 και 12 (10 και 17 Δεκεμβρίου) το μάθημα γίνεται εργαστήριο πάνω στην εργασία, με τον διδάσκοντα παρόντα: την πρώτη φορά δουλεύουμε το Μέρος Α, τη δεύτερη το Μέρος Β. Φέρτε το notebook σας και τις ερωτήσεις σας. Εκτός μαθήματος, ρωτήστε στο φόρουμ του eClass (ώστε να δουν την απάντηση όλοι) ή στις ώρες γραφείου. Για σφάλματα Python: πρώτα διαβάστε το μήνυμα (ξέρετε πια τα βασικά), μετά τη σελίδα της αντίστοιχης εβδομάδας, μετά ρωτήστε, μαζί με το μήνυμα και τον κώδικα που το προκάλεσε.

ΑΚΑΔΗΜΑΪΚΗ ΑΚΕΡΑΙΟΤΗΤΑ

Μπορείτε να συμβουλευτείτε οποιαδήποτε πηγή, συμπεριλαμβανομένων εργαλείων AI, για να *μάθετε*. Το notebook όμως πρέπει να είναι δικό σας και να μπορείτε να εξηγήσετε κάθε γραμμή του στην παρουσίαση. Κώδικας που δεν μπορείτε να εξηγήσετε δεν βαθμολογείται. Αναφέρετε στο τέλος τι πηγές ή εργαλεία χρησιμοποιήσατε και πώς.

ΚΑΛΗ ΔΟΥΛΕΙΑ

Η πρώτη φορά που θα δείτε τα 32 εκατομμύρια γραμμές να απαντούν σε μία ερώτηση με μία γραμμή groupby είναι η στιγμή που καταλαβαίνετε γιατί μάθατε όλα αυτά. Απολαύστε την.

Ευάγγελος Καλαμπόκης | Information Systems Lab | Πανεπιστήμιο Μακεδονίας | Δεδομένα: The Instacart Online Grocery Shopping Dataset 2017. Ο κώδικας του μίνι notebook εκτελείται στον browser με Pyodide.