



ΕΑΠ

ΜΕΤΑΠΤΥΧΙΑΚΗ ΕΞΕΙΔΙΚΕΥΣΗ ΣΤΑ
ΠΛΗΡΟΦΟΡΙΑΚΑ ΣΥΣΤΗΜΑΤΑ

Διπλωματική Εργασία

Αναλυτική Επεξεργασία Δεδομένων των Αθλητικών
Διοργανώσεων
“World Marathon Majors”

Πάριος Κάτσης

Επιβλέπων καθηγητής: Μιχαήλ Βασιλακόπουλος

Πάτρα, Ιούνιος 2026

Η παρούσα εργασία αποτελεί πνευματική ιδιοκτησία του φοιτητή («συγγραφέας/δημιουργός») που την εκπόνησε. Στο πλαίσιο της πολιτικής ανοικτής πρόσβασης ο συγγραφέας/δημιουργός εκχωρεί στο ΕΑΠ, μη αποκλειστική άδεια χρήσης του δικαιώματος αναπαραγωγής, προσαρμογής, δημόσιου δανεισμού, παρουσίασης στο κοινό και ψηφιακής διάχυσής τους διεθνώς, σε ηλεκτρονική μορφή και σε οποιοδήποτε μέσο, για διδακτικούς και ερευνητικούς σκοπούς, άνευ ανταλλάγματος και για όλο το χρόνο διάρκειας των δικαιωμάτων πνευματικής ιδιοκτησίας. Η ανοικτή πρόσβαση στο πλήρες κείμενο για μελέτη και ανάγνωση δεν σημαίνει καθ' οιονδήποτε τρόπο παραχώρηση δικαιωμάτων διανοητικής ιδιοκτησίας του συγγραφέα/δημιουργού ούτε διευκολύνει την αναπαραγωγή, αναδημοσίευση, αντιγραφή, αποθήκευση, πώληση, εμπορική χρήση, μετάδοση, διανομή, έκδοση, εκτέλεση, «μεταφόρτωση» (downloading), «ανάρτηση» (uploading), μετάφραση, τροποποίηση με οποιονδήποτε τρόπο, τμηματικά ή περιληπτικά της εργασίας, χωρίς τη ρητή προηγούμενη έγγραφη συναίνεση του συγγραφέα/δημιουργού. Ο συγγραφέας/δημιουργός διατηρεί το σύνολο των ηθικών και περιουσιακών του δικαιωμάτων.



Αναλυτική Επεξεργασία Δεδομένων των Αθλητικών
Διοργανώσεων
“World Marathon Majors”

Πάριος Κάτσης

Επιτροπή Επίβλεψης Διπλωματικής Εργασίας

Επιβλέπων Καθηγητής:
Μιχαήλ Βασιλακόπουλος
Καθηγητής Παν. Θεσσαλίας

Συν-Επιβλέπων Καθηγητής:
Γεώργιος Μαυρομάτης
Μέλος ΣΕΠ ΕΑΠ

Πάτρα, Ιούνιος 2026

Θα ήθελα να εκφράσω τις ειλικρινείς μου ευχαριστίες στον επιβλέποντα καθηγητή μου, κ. Μιχαήλ Βασιλακόπουλο, για την καθοδήγηση και την υποστήριξή του καθ' όλη τη διάρκεια εκπόνησης της παρούσας εργασίας. Επίσης, ευχαριστώ θερμά τον συν-επιβλέποντα καθηγητή, κ. Γεώργιο Μαυρομάτη, για τις πολύτιμες παρατηρήσεις του. Η εργασία αυτή αφιερώνεται στην οικογένειά μου για την αμέριστη συμπαράστασή της.

Περίληψη

Η ανάλυση αθλητικών δεδομένων συνιστά έναν κλάδο της επιστήμης δεδομένων που επικεντρώνεται στη συλλογή, επεξεργασία, ανάλυση και ερμηνεία δεδομένων που προέρχονται από αθλητικά περιβάλλοντα. Χρησιμοποιεί στατιστικές μεθόδους, μηχανική μάθηση και τεχνικές διαχείρισης Big Data για την εξαγωγή χρήσιμων συμπερασμάτων. Ο βασικός της στόχος είναι η βαθύτερη κατανόηση της ανθρώπινης φυσιολογίας σε συνθήκες ακραίας καταπόνησης, η βελτίωση της αγωνιστικής απόδοσης, η πρόληψη της πρόωρης κόπωσης και η λήψη τεκμηριωμένων αποφάσεων σχετικά με τη στρατηγική ρυθμού (pacing) στα αγωνίσματα αντοχής.

Οι εξελίξεις στην τεχνολογία, η χρήση αθλητικών ρολογιών με GPS και η ψηφιοποίηση των αποτελεσμάτων έχουν διευκολύνει τη συλλογή αυτής της πληροφορίας σε μαζική κλίμακα. Οι διοργανώσεις καταγράφουν πλέον με ακρίβεια δευτερολέπτου τις επιδόσεις χιλιάδων δρομέων, παρέχοντας δεδομένα για τον καθαρό χρόνο τερματισμού, τα ενδιάμεσα περάσματα (splits) και τον ρυθμό τρεξίματος, τα οποία σε συνδυασμό με ιστορικά μετεωρολογικά δεδομένα επιτρέπουν τη μελέτη της επίδρασης του εξωτερικού περιβάλλοντος στον ανθρώπινο οργανισμό.

Η εργασία αυτή στοχεύει στην ανάλυση ιστορικών δεδομένων από τους έξι κορυφαίους διεθνείς μαραθωνίους (World Marathon Majors) και τον Αυθεντικό Μαραθώνιο της Αθήνας, με σκοπό τον ποσοτικό προσδιορισμό της επίδρασης της περιβαλλοντικής θερμοκρασίας και του υψομετρικού αναγλύφου στη στρατηγική pacing και την τελική επίδοση των αθλητών. Αρχικά έγινε η ανεύρεση και εξαγωγή των δεδομένων τερματισμάτων δρομέων μέσω τεχνικών Web Scraping με τη χρήση της γλώσσας προγραμματισμού Python. Έπειτα, ακολούθησε η επεξεργασία, ο καθαρισμός και η κατάλληλη μορφοποίησή τους, ώστε να εισαχθούν δομημένα σε μια σχεσιακή βάση δεδομένων SQL εντός του περιβάλλοντος SQL Server Management Studio (SSMS). Μετά τη δόμηση του πολυδιάστατου μοντέλου δεδομένων, διεξήχθη live σύνδεση με ένα εργαλείο επιχειρηματικής ευφυΐας και οπτικοποίησης για τη δημιουργία διαδραστικών πινάκων ελέγχου (dashboards). Τέλος, εφαρμόστηκαν επαγωγικοί στατιστικοί έλεγχοι (ANOVA, t-test, Chi-Square), αναδεικνύοντας ότι οι υψηλές θερμοκρασίες υποβαθμίζουν σημαντικά την απόδοση, δεν παρατηρήθηκε στατιστικά σημαντική διαφορά στη στρατηγική pacing μεταξύ ανδρών και γυναικών ($p = 0,323$), ενώ το έντονο υψομετρικό

προφίλ της Αθήνας αποτρέπει συστηματικά την επίτευξη αρνητικού περάσματος (negative split).

Λέξεις – Κλειδιά

Αναλυτική αθλητικών δεδομένων, επιχειρηματική ευφυΐα, στρατηγική ρυθμού, μααραθώνιος δρόμος, θερμικό στρες, υψομετρικό ανάγλυφο.

Analytical Processing of Sports Event Data

“World Marathon Majors”

Master in Information Systems

Parios Katsis

Abstract

Sports data analytics is a branch of data science that focuses on the collection, processing, analysis, and interpretation of data derived from athletic environments. It utilizes statistical methods, machine learning, and Big Data management techniques to extract valuable insights. Its primary objective is to deepen the understanding of human physiology under extreme physical exertion, optimize athletic performance, prevent premature fatigue, and support data-driven decisions regarding pacing strategies in endurance sports.

Advancements in technology, the widespread use of GPS sports watches, and the digitization of race results have significantly facilitated the collection of this information on a massive scale. Modern running events accurately record the performance of thousands of runners to the exact second, providing structured data on net finish times, intermediate splits, and running pace, which, when combined with historical meteorological data, enable the comprehensive study of environmental impacts on the human body.

This thesis aims to analyze historical data from the six World Marathon Majors and the Authentic Athens Marathon to quantitatively determine the impact of ambient temperature and course topography on athletes' pacing strategies and final performance. Initially, data extraction of finishing runners was achieved through Web Scraping techniques using the Python programming language. Subsequently, data processing, cleaning, and standardization were performed to load the information into a structured relational SQL database within SQL Server Management Studio (SSMS). Following the development of the multidimensional data model, a live connection was established with a business intelligence and visualization tool to construct interactive dashboards. Finally, inferential statistical tests (ANOVA, t-tests, Chi-Square) were applied, demonstrating that elevated

temperatures significantly degrade running performance, no statistically significant difference in pacing strategy was found between male and female runners (H_2 rejected, $p = 0.323$), and the demanding elevation profile of the Athens course systematically prevents the execution of a negative split strategy.

Keywords

Sports data analytics, business intelligence, pacing strategy, marathon running, thermal stress, course topography.

Περιεχόμενα

Περίληψη	vi
Abstract	viii
Περιεχόμενα.....	x
Κατάλογος Εικόνων / Σχημάτων	xii
Κατάλογος Πινάκων	xiii
Συντομογραφίες & Ακρωνύμια	xiv
1. Εισαγωγή.....	1
1.1 Παρουσίαση του Θέματος	1
1.1.1 Το Φαινόμενο του Σύγχρονου Μαραθωνίου	2
1.1.2 Ψηφιοποίηση και Δεδομένα Αγώνων	2
1.1.3 Η Τεχνολογική Υποδομή της Έρευνας.....	2
1.2 Σημαντικότητα της Έρευνας.....	3
1.2.1 Χάσμα Γνώσης και Ερευνητική Αναγκαιότητα	3
1.2.2 Επιστημονική και Πρακτική Σημασία	4
1.2.3 Συμβολή στην Αθλητική Αναλυτική	4
1.3 Σκοπός και Αντικείμενο της Διπλωματικής Εργασίας.....	4
1.4 Ερευνητικά Ερωτήματα και Υποθέσεις.....	5
1.5 Δομή Εργασίας	5
2. Ανασκόπηση Βιβλιογραφίας.....	7
2.1 Φυσιολογία του Τρεξίματος Αντοχής και Θερμορρύθμιση	7
2.1.1 Φυσιολογικές Βάσεις Θερμικής Καταπόνησης	8
2.1.2 Κεντρικός Ρυθμιστής και Αντιληπτή Κόπωση	9
2.2 Στρατηγικές Ρυθμού (Pacing Strategies) σε Δρόμους Αντοχής	9
2.3 Επίδραση Περιβαλλοντικών Παραγόντων στην Απόδοση	10
2.4 Γεωγραφικό Ανάγλυφο και Μηχανική του Τρεξίματος.....	10
3. Μεθοδολογία	13
3.1 Ερευνητικός Σχεδιασμός	13
3.2 Πληθυσμός και Δείγμα της Έρευνας.....	14
3.3 Επιχειρησιακός Ορισμός και Κατηγοριοποίηση Μεταβλητών.....	15
3.3.1 Ανεξάρτητες Μεταβλητές (Προβλεπτικοί Παράγοντες)	15
3.3.2 Εξαρτημένες Μεταβλητές (Μεταβλητές Έκβασης)	16
3.4 Αρχιτεκτονική Πληροφοριακού Συστήματος και Διαχείριση Δεδομένων	17
3.5 Στατιστικό Πλάνο Ανάλυσης και Έλεγχος Υποθέσεων.....	18
3.6 Προϋποθέσεις Στατιστικών Ελέγχων	19
4. Υποδομή Αναλυτικής Επεξεργασίας Δεδομένων	21
4.1 Sports Data Analytics στον Αγώνα Αντοχής.....	21
4.2 Αποθήκη Δεδομένων (Data Warehouse).....	22
4.2.1 Η Διαδικασία ETL (Extract, Transform, Load)	23
4.2.2 Πολυδιάστατο Μοντέλο Δεδομένων	23
4.3 On-Line Analytical Processing (OLAP) και Συναθροίσεις	25
4.4 Τύποι Συστημάτων OLAP	28
5. Τεχνολογικά Εργαλεία και Περιβάλλοντα Ανάπτυξης.....	31
5.1 Γλώσσα Προγραμματισμού Python	31

5.1.1 Το Περιβάλλον Ανάπτυξης Spyder	31
5.2 Microsoft SQL Server	32
5.2.1 Εκδόσεις του Microsoft SQL Server	33
5.2.2 Δομή και Διασύνδεση των Δεδομένων	33
5.3 Microsoft Power BI	34
5.4 Ανάπτυξη Διαδραστικού Dashboard στο Power BI	34
6. Συλλογή Δεδομένων	39
6.1 Διεθνείς Βάσεις Δεδομένων Μαραθωνίων	39
6.1.1 Η Τεχνική Web Scraping	39
6.2 Μετεωρολογικές Βάσεις Δεδομένων (Open Climate Data)	40
6.2.1 Διασύνδεση με Ανοικτά Δεδομένα Κλίματος	41
6.2.2 Δομή και Διασύνδεση των Δεδομένων	41
7. Αποτελέσματα	43
7.1 Περιγραφική Στατιστική Ανάλυση Δείγματος	43
7.2 Έλεγχος 1ης Υπόθεσης (Θερμοκρασία και Απόδοση)	45
7.3 Έλεγχος 2ης Υπόθεσης (Φύλο και Κατανομή Ρυθμού)	47
7.4 Έλεγχος 3ης Υπόθεσης (Υψομετρικό Ανάγλυφο και Στρατηγική Splits)	49
8. Συζήτηση	53
8.1 Ερμηνεία των Ευρημάτων και Φυσιολογική Μηχανική	53
8.2 Σύγκριση με τη Διεθνή Βιβλιογραφία	54
8.3 Θεωρητικές και Πρακτικές Προεκτάσεις	54
9. Συμπεράσματα	57
9.1 Σύνοψη Κύριων Ευρημάτων	57
9.2 Περιορισμοί της Έρευνας	58
9.3 Προτάσεις για Μελλοντική Έρευνα	58
Βιβλιογραφία	61
Παράρτημα Β: Κώδικας Στατιστικής Ανάλυσης	67

Κατάλογος Εικόνων / Σχημάτων

Εικόνα 1 Sports Analytics	7
Εικόνα 2 Αρχιτεκτονική Αποθήκης Δεδομένων	14
Εικόνα 3 Σχήμα Αστέρα (Star Schema) της Αποθήκης Δεδομένων	25
Εικόνα 4 OLAP Cube — Πολυδιάστατη Αναπαράσταση Δεδομένων	26
Εικόνα 5 Στιγμιότυπο διαδραστικού πίνακα ελέγχου (Power BI dashboard) – Συμμετοχές ανά Μαραθώνιο και φύλο	45
Εικόνα 6 Στιγμιότυπο διαδραστικού πίνακα ελέγχου (Power BI dashboard) – Μέσος χρόνος ανά κατηγορία θερμοκρασίας.....	47
Εικόνα 7 Στιγμιότυπο διαδραστικού πίνακα ελέγχου (Power BI dashboard) – Μέσος χρόνος ανά φύλο	49
Εικόνα 8 Στιγμιότυπο διαδραστικού πίνακα ελέγχου (Power BI dashboard) – Αρνητικά splits ανά κατηγορία υψομέτρου	51

Κατάλογος Πινάκων

Πίνακας 1 Περιγραφική Στατιστική του Δείγματος (N=105.000)	44
Πίνακας 2 Αποτελέσματα One-Way ANOVA — Επίδραση Θερμοκρασίας στον Χρόνο Τερματισμού	46
Πίνακας 3 Αποτελέσματα Independent Samples t-test — Φύλο και Pacing Ratio	48
Πίνακας 4 Αποτελέσματα Ελέγχου χ^2 — Υψομετρικό Ανάγλυφο και Αρνητικό Split	50

Συντομογραφίες & Ακρωνύμια

ΔΕ	Διπλωματική Εργασία
ΕΑΠ	Ελληνικό Ανοικτό Πανεπιστήμιο
ΕΕ	Ερευνητικό Ερώτημα
ΕΥ	Ερευνητική Υπόθεση
ΤΠΕ	Τεχνολογίες Πληροφορικής και Επικοινωνιών
WMM	World Marathon Majors
API	Application Programming Interface (Διεπαφή Προγραμματισμού Εφαρμογών)
BI	Business Intelligence (Επιχειρηματική Ευφυΐα)
CSV	Comma-Separated Values (Αρχείο τιμών διαχωρισμένων με κόμμα)
ETL	Extract, Transform, Load (Εξαγωγή, Μετασχηματισμός, Φόρτωση)
GPS	Global Positioning System (Παγκόσμιο Σύστημα Εντοπισμού Θέσης)
HOLAP	Hybrid Online Analytical Processing
HTML	HyperText Markup Language
IDE	Integrated Development Environment (Ολοκληρωμένο Περιβάλλον Ανάπτυξης)
MOLAP	Multidimensional Online Analytical Processing
OLAP	On-Line Analytical Processing (Αναλυτική Επεξεργασία Άμεσης Απόκρισης)
RDBMS	Relational Database Management System (Σχεσιακό Σύστημα Διαχείρισης Βάσεων Δεδομένων)
ROLAP	Relational Online Analytical Processing
SQL	Structured Query Language (Δομημένη Γλώσσα Ερωτημάτων)
SSMS	SQL Server Management Studio
URL	Uniform Resource Locator

1. Εισαγωγή

Το παρόν κεφάλαιο θέτει τα θεμέλια της διπλωματικής εργασίας παρουσιάζοντας το ερευνητικό πλαίσιο και τη σημαντικότητα της μελέτης. Αναπτύσσονται ο σκοπός και τα ερευνητικά ερωτήματα που κατευθύνουν τη μελέτη, ενώ στο τέλος του κεφαλαίου περιγράφεται η δομή της εργασίας, ώστε ο αναγνώστης να έχει σαφή εικόνα της ακολουθίας που θα ακολουθήσει.

1.1 Παρουσίαση του Θέματος

Η ραγδαία ανάπτυξη των Τεχνολογιών Πληροφορικής και Επικοινωνιών (ΤΠΕ) κατά την τελευταία εικοσαετία έχει μεταμορφώσει ριζικά τον τομέα του αθλητισμού, εισάγοντας την επιστήμη των δεδομένων στην καθημερινή προπονητική και αγωνιστική πρακτική. Το πεδίο της Ανάλυσης Αθλητικών Δεδομένων (Sports Data Analytics) έχει καταστεί πλέον αυτόνομος και συνεχώς εξελισσόμενο κλάδο της Αθλητικής Επιστήμης (Sport Science), ο οποίος συνδυάζει τη στατιστική μοντελοποίηση, τη μηχανική μάθηση και τη διαχείριση μεγάλων συνόλων δεδομένων (Big Data) με σκοπό τη βελτιστοποίηση της αθλητικής απόδοσης, την πρόληψη τραυματισμών και τη στρατηγική λήψη αποφάσεων.

Στο πλαίσιο των αγωνισμάτων αντοχής, και ειδικότερα του Μαραθωνίου δρόμου (42.195 μέτρα), η συλλογή και η ανάλυση δεδομένων παρουσιάζουν ιδιαίτερο ενδιαφέρον, καθώς ο Μαραθώνιος δεν αντιπροσωπεύει μόνο μια δοκιμασία των φυσιολογικών ορίων του ανθρώπινου οργανισμού, αλλά και ένα σύνθετο πολυπαραγοντικό πρόβλημα. Η τελική επίδοση επηρεάζεται από τη στρατηγική κατανομής ρυθμού (pacing), τα δημογραφικά χαρακτηριστικά του δρομέα, καθώς και από εξωτερικές μεταβλητές. Η διεθνής κοινότητα εστιάζει την προσοχή της στους κορυφαίους αγώνες του πλανήτη, γνωστούς ως **World Marathon Majors** (Τόκιο, Βοστώνη, Λονδίνο, Βερολίνο, Σικάγο, Νέα Υόρκη), οι οποίοι προσελκύουν εκατοντάδες χιλιάδες δρομείς υψηλού επιπέδου και ερασιτέχνες, παράγοντας έναν τεράστιο όγκο πρωτογενών δεδομένων.

1.1.1 Το Φαινόμενο του Σύγχρονου Μαραθωνίου

Ο σύγχρονος Μαραθώνιος δρόμος ανήκει στα πιο δημοφιλή αθλητικά φαινόμενα παγκοσμίως. Από τα 300.000 τερματισμούς που καταγράφηκαν παγκοσμίως το 1990, ο αριθμός ξεπέρασε τα 1,1 εκατομμύρια ετησίως τη δεκαετία του 2010, αντανakλώντας μια εκθετική ανάπτυξη της μαζικής συμμετοχής. Οι έξι κορυφαίοι μαραθώνιοι του κόσμου (World Marathon Majors) - Τόκιο, Βοστώνη, Λονδίνο, Βερολίνο, Σικάγο και Νέα Υόρκη - προσελκύουν από κοινού άνω των 250.000 δρομέων ετησίως, δημιουργώντας τεράστιους όγκους αγωνιστικών δεδομένων που συνιστούν πολύτιμη πηγή επιστημονικής έρευνας και αθλητικής αναλυτικής.

1.1.2 Ψηφιοποίηση και Δεδομένα Αγώνων

Η ψηφιοποίηση των αγωνιστικών αποτελεσμάτων και η υιοθέτηση ηλεκτρονικής χρονομέτρησης με τσιπ RFID έχουν μεταμορφώσει τον τρόπο συλλογής δεδομένων στους μαραθωνίους. Σήμερα, κάθε δρομέας καταγράφεται με ακρίβεια δευτερολέπτου σε πολλαπλά σημεία της διαδρομής, παρέχοντας δεδομένα για τον καθαρό χρόνο τερματισμού (Net Time), τον χρόνο διέλευσης στο μισό (Half Marathon Time) και τον ρυθμό τρεξίματος ανά χιλιόμετρο. Τα επίσημα αποτελέσματα δημοσιεύονται σε ψηφιακές πλατφόρμες, καθιστώντας εφικτή την αυτοματοποιημένη συλλογή τους μέσω τεχνικών Web Scraping, όπως αυτές που εφαρμόστηκαν στην παρούσα εργασία.

1.1.3 Η Τεχνολογική Υποδομή της Έρευνας

Η τεχνολογική υποδομή που χρησιμοποιήθηκε στην παρούσα εργασία βασίζεται σε τρία αλληλένδετα εργαλεία: (α) τη γλώσσα προγραμματισμού Python για τη συλλογή και επεξεργασία των δεδομένων μέσω Web Scraping και της βιβλιοθήκης Pandas, (β) τον Microsoft SQL Server ως σχεσιακό σύστημα διαχείρισης βάσεων δεδομένων για τη δομημένη αποθήκευση των 105.000 εγγράφων σε σχεσιακό σχήμα αστέρα (πίνακες γεγονότων και πίνακες διαστάσεων), και (γ) το Microsoft Power BI για τη δημιουργία διαδραστικών dashboards και τη γεφύρωση της τεχνικής ανάλυσης με την οπτικοποίηση αθλητικών τάσεων. Η αρχιτεκτονική αυτή συνιστά ένα ολοκληρωμένο σύστημα Επιχειρηματικής Ευφυΐας (Business Intelligence) εφαρμοσμένο στον τομέα της αθλητικής αναλυτικής.

1.2 Σημαντικότητα της Έρευνας

Η σημασία της παρούσας διπλωματικής εργασίας έγκειται στην ανάγκη γεφύρωσης του χάσματος μεταξύ της θεωρητικής αθλητικής φυσιολογίας και της πρακτικής εφαρμογής της αναλυτικής δεδομένων σε πραγματικές συνθήκες αγώνα. Ενώ οι εργαστηριακές μετρήσεις προσφέρουν πολύτιμες πληροφορίες για τη μέγιστη πρόσληψη οξυγόνου (VO_{2max}) και το αναερόβιο κατώφλι, δεν μπορούν να ενσωματώσουν πλήρως τη δυναμική αλληλεπίδραση των κλιματικών αλλαγών και του γεωγραφικού αναγλύφου μιας διαδρομής.

Μέσω της χρήσης σύγχρονων εργαλείων πληροφορικής, όπως η γλώσσα προγραμματισμού Python για τη συλλογή δεδομένων (Web Scraping), οι σχεσιακές βάσεις δεδομένων (SQL Server) για τη δόμηση και οργάνωση της πληροφορίας, και τα συστήματα επιχειρηματικής ευφυΐας (Power BI) για την οπτικοποίηση, η έρευνα αυτή διευκολύνει την ταυτόχρονη εξέταση εκατοντάδων εγγραφών. Η συμπερίληψη του Αυθεντικού Μαραθωνίου Αθηνών δίπλα στους World Marathon Majors προσφέρει μια μοναδική ευκαιρία σύγκρισης μιας ιστορικής, ιδιαίτερα απαιτητικής και υψομετρικά έντονης διαδρομής με τις παραδοσιακά "γρήγορες" και επίπεδες διαδρομές του εξωτερικού, όπως του Βερολίνου και του Σικάγο. Τα συμπεράσματα που προκύπτουν έχουν άμεση εφαρμογή στον σχεδιασμό εξατομικευμένων πλάνων ρυθμού για αθλητές και προπονητές.

1.2.1 Χάσμα Γνώσης και Ερευνητική Αναγκαιότητα

Παρά το πλούσιο βιβλιογραφικό υπόβαθρο στη φυσιολογία του μαραθωνίου, παρατηρείται ένα σημαντικό ερευνητικό κενό: η συστηματική, ποσοτική σύγκριση της επίδρασης της θερμοκρασίας και του υψομετρικού αναγλύφου σε πραγματικά δεδομένα μεγάλης κλίμακας από πολλαπλές διοργανώσεις ταυτόχρονα. Οι περισσότερες μελέτες εστιάζουν σε έναν αγώνα ή χρησιμοποιούν μικρά δείγματα εργαστηριακών μετρήσεων. Η παρούσα έρευνα αξιοποιεί σύγχρονα εργαλεία πληροφορικής για να γεφυρώσει αυτό το χάσμα, αναλύοντας 105.000 αγωνιστικές εγγραφές από επτά διαφορετικές διοργανώσεις με ανομοιογενή κλιματικά και τοπογραφικά χαρακτηριστικά.

1.2.2 Επιστημονική και Πρακτική Σημασία

Η επιστημονική αξία της παρούσας έρευνας έγκειται στην εφαρμογή ποσοτικών στατιστικών μεθόδων σε μεγάλο σύνολο πραγματικών αγωνιστικών δεδομένων. Τα ευρήματα μπορούν να χρησιμοποιηθούν από προπονητές, αθλητές και διοργανωτές για τη βελτίωση αγωνιστικών στρατηγικών.

1.2.3 Συμβολή στην Αθλητική Αναλυτική

Η εργασία συμβάλλει στο πεδίο της αθλητικής αναλυτικής με την ανάπτυξη ολοκληρωμένης αρχιτεκτονικής πληροφοριακού συστήματος (Python → SQL Server → Power BI) για τη διαχείριση και οπτικοποίηση αθλητικών δεδομένων μαραθωνίου.

1.3 Σκοπός και Αντικείμενο της Διπλωματικής Εργασίας

Ο κύριος σκοπός της παρούσας διπλωματικής εργασίας είναι η συστηματική ανάλυση ιστορικών δεδομένων από τους World Marathon Majors και τον Μαραθώνιο Αθηνών, με στόχο τον ποσοτικό προσδιορισμό της επίδρασης του εξωτερικού περιβάλλοντος (θερμοκρασία) και της μορφολογίας της διαδρομής (υψομετρικά) στην αγωνιστική συμπεριφορά και την τελική επίδοση των δρομέων. Επιπλέον, εξετάζεται η διαφοροποίηση των στρατηγικών racing μεταξύ των δύο φύλων, προκειμένου να διαπιστωθεί εάν υφίστανται σταθερά πρότυπα διαχείρισης δυνάμεων κατά τη διάρκεια του αγώνα.

Το αντικείμενο της έρευνας περιορίζεται στην επεξεργασία του δομημένου συνόλου δεδομένων All_Marathons.csv, το οποίο εμπεριέχει καταγεγραμμένες επιδόσεις αθλητών που τερμάτισαν επιτυχώς. Μέσα από τη στατιστική επεξεργασία αυτών των δεδομένων, επιδιώκεται η δημιουργία ενός προγνωστικού και περιγραφικού πλαισίου που θα αναδεικνύει τις βέλτιστες συνθήκες για την επίτευξη ατομικών ρεκόρ (Personal Bests) και θα τροφοδοτεί διαδραστικά dashboards ψηφιακής αναλυτικής.

1.4 Ερευνητικά Ερωτήματα και Υποθέσεις

Για την επίτευξη του ανωτέρω σκοπού, η παρούσα μελέτη διαρθρώνεται γύρω από τρία κεντρικά Ερευνητικά Ερωτήματα (ΕΕ), καθένα από τα οποία συνδέεται με μια αντίστοιχη Ερευνητική Υπόθεση (Η):

- **Ερευνητικό Ερώτημα 1 (ΕΕ1):** Πώς επηρεάζει η περιβαλλοντική θερμοκρασία (όπως αυτή κατηγοριοποιείται στη μεταβλητή Temp_Category) τον τελικό χρόνο ολοκλήρωσης (Net_Time) και τον προσαρμοσμένο δείκτη απόδοσης (Performance_Index) των δρομέων;
 - *Ερευνητική Υπόθεση 1 (H1):* Οι υψηλές περιβαλλοντικές θερμοκρασίες σχετίζονται με στατιστικά σημαντική υποβάθμιση του δείκτη απόδοσης και αύξηση του συνολικού χρόνου ολοκλήρωσης του αγώνα, λόγω της θερμικής καταπόνησης.
- **Ερευνητικό Ερώτημα 2 (ΕΕ2):** Υπάρχει στατιστικά σημαντική διαφοροποίηση στη στρατηγική κατανομής ρυθμού (όπως εκφράζεται από τη μεταβλητή Pacing_ratio και την επίτευξη Negative_split) μεταξύ ανδρών και γυναικών δρομέων;
 - *Ερευνητική Υπόθεση 2 (H2):* Οι γυναίκες δρομείς επιδεικνύουν πιο σταθερή και ορθολογική στρατηγική pacing (τιμή Pacing_ratio εγγύτερα στη μονάδα) σε σχέση με τους άνδρες, οι οποίοι παρουσιάζουν μεγαλύτερη επιβράδυνση στο δεύτερο μισό του αγώνα.
- **Ερευνητικό Ερώτημα 3 (ΕΕ3):** Σε ποιο βαθμό η υψομετρική δυσκολία και η μορφολογία της διαδρομής (Elevation_Category) αποτρέπει ή διευκολύνει την επίτευξη αρνητικού περάσματος (negative split) κατά τη διάρκεια του μααραθωνίου;
 - *Ερευνητική Υπόθεση 3 (H3):* Διαδρομές με έντονο υψομετρικό προφίλ (κατηγορία Moderate/Hilly), όπως ο Μαραθώνιος Αθηνών, παρουσιάζουν στατιστικά σημαντική μείωση της πιθανότητας επίτευξης αρνητικού περάσματος σε σύγκριση με επίπεδες διαδρομές.

1.5 Δομή Εργασίας

Η παρούσα διπλωματική εργασία διαρθρώνεται σε εννέα (9) κεφάλαια. Το Κεφάλαιο 1 παρουσιάζει το θέμα, τους σκοπούς και τις ερευνητικές υποθέσεις. Το Κεφάλαιο 2

αφιερώνεται στην ανασκόπηση βιβλιογραφίας στη φυσιολογία του μαραθωνίου και τις στρατηγικές racing. Το Κεφάλαιο 3 περιγράφει τη μεθοδολογία και τον ερευνητικό σχεδιασμό. Το Κεφάλαιο 4 αναπτύσσει τις θεωρητικές έννοιες των τεχνολογιών (Data Warehouse, OLAP, Star Schema). Το Κεφάλαιο 5 περιγράφει τα τεχνολογικά εργαλεία (Python, SQL Server, Power BI). Το Κεφάλαιο 6 παρουσιάζει τη συλλογή και προεπεξεργασία δεδομένων. Το Κεφάλαιο 7 παρουσιάζει τα αποτελέσματα των στατιστικών ελέγχων. Το Κεφάλαιο 8 περιέχει τη σύγκριση με τη βιβλιογραφία και τη συζήτηση. Το Κεφάλαιο 9 παρουσιάζει τα συμπεράσματα και τις προτάσεις για μελλοντική έρευνα.

2. Ανασκόπηση Βιβλιογραφίας

Το παρόν κεφάλαιο συνοψίζει την τρέχουσα επιστημονική γνώση στους βασικούς θεματικούς άξονες της έρευνας. Εξετάζεται η φυσιολογία του τρεξίματος αντοχής και η θερμορρύθμιση, οι στρατηγικές κατανομής ρυθμού (pacing strategies), η επίδραση των περιβαλλοντικών παραγόντων στην αθλητική απόδοση, καθώς και ο ρόλος του γεωγραφικού αναγλύφου στη μηχανική του τρεξίματος.

2.1 Φυσιολογία του Τρεξίματος Αντοχής και Θερμορρύθμιση

Στην Εικόνα 1 παρουσιάζεται το εννοιολογικό πλαίσιο της αναλυτικής αθλητικών δεδομένων (Sports Data Analytics), αποτυπώνοντας σχηματικά τις τρεις κεντρικές φάσεις: συλλογή, ανάλυση και οπτικοποίηση αγωνιστικών δεδομένων. Το διάγραμμα αναδεικνύει τη συστηματική ροή πληροφορίας από την πρωτογενή πηγή - επίσημα χρονομετρικά αρχεία και βιομετρικοί αισθητήρες - έως την παραγωγή πρακτικής γνώσης για αθλητές και προπονητές. Το κεντρικό συμπέρασμα που αναδύεται είναι ότι η αθλητική αναλυτική δεν συνιστά ένα μεμονωμένο εργαλείο, αλλά ένα ολοκληρωμένο σύστημα διαδοχικών διεργασιών που μετατρέπει ακατέργαστα αγωνιστικά δεδομένα σε επιχειρήσιμες πληροφορίες υποστήριξης αποφάσεων.

Η ολοκλήρωση του μαραθωνίου δρόμου απαιτεί τη συνεχή διατήρηση ενός υψηλού ποσοστού της μέγιστης πρόσληψης οξυγόνου ($\text{VO}_{2\text{max}}$) για παρατεταμένο χρονικό διάστημα, γεγονός που συνεπάγεται ακραία μεταβολική και θερμική επιβάρυνση του ανθρώπινου οργανισμού (Joyner and Coyle, 2008; Κοτζαμανίδης, 2019). Κατά τη διάρκεια της παρατεταμένης μυϊκής σύσπασης, η μηχανική απόδοση του σκελετικού μυϊκού



Εικόνα 1 Sports Analytics

συστήματος κυμαίνεται μεταξύ 20% και 25%, πράγμα που σημαίνει ότι το εναπομείναν 75%–80% της καταναλισκόμενης ενέργειας απελευθερώνεται υπό τη μορφή μεταβολικής θερμότητας. Η εσωτερική θερμοκρασία του πυρήνα του σώματος αυξάνεται αναλογικά με την ένταση και τη διάρκεια της άσκησης, ενεργοποιώντας άμεσα τους ομοιοστατικούς μηχανισμούς θερμορρύθμισης που ελέγχονται από τον υποθάλαμο.

Όταν η παραγωγή μεταβολικής θερμότητας υπερβαίνει τον ρυθμό αποβολής της στο εξωτερικό περιβάλλον, ο οργανισμός επιστρατεύει τη δερματική αγγειοδιαστολή και την έκκριση ιδρώτα ως κύριους μηχανισμούς ψύξης μέσω εξάτμισης. Ωστόσο, η αυξημένη ανακατανομή του αίματος προς την περιφέρεια (δέρμα) για τη διευκόλυνση της θερμικής αποβολής προκαλεί ανταγωνισμό με τις ανάγκες των εργαζόμενων μυών για οξυγόνο και θρεπτικά συστατικά. Αυτή η κατάσταση περιορίζει τη φλεβική επιστροφή και μειώνει τον όγκο παλμού της καρδιάς. Για τη διατήρηση του καρδιακού δείκτη (cardiac output), παρατηρείται μια προοδευτική αύξηση της καρδιακής συχνότητας, ένα φαινόμενο γνωστό ως καρδιαγγειακή μετατόπιση (cardiovascular drift) (Coyle and González-Alonso, 2001).

Σύμφωνα με το Μοντέλο του Κεντρικού Ρυθμιστή (Central Governor Model), η μείωση της αθλητικής απόδοσης και η εμφάνιση της κόπωσης δεν οφείλονται αποκλειστικά σε περιφερειακά καταστροφικά σφάλματα των μυών (όπως η εξάντληση του μυϊκού γλυκογόνου ή η συσσώρευση μεταβολιτών), αλλά αποτελούν μια εγκεφαλικά ελεγχόμενη, προστατευτική σύσπαση της κινητικής επιστράτευσης. Το κεντρικό νευρικό σύστημα ρυθμίζει υποσυνείδητα τον ρυθμό τρεξίματος, αξιολογώντας συνεχώς τα προσαγωγά μηνύματα από τους περιφερειακούς χημειοϋποδοχείς και θερμοϋποδοχείς. Στόχος της συγκεκριμένης ρύθμισης είναι η αποτροπή της ανόδου της εσωτερικής θερμοκρασίας σε κρίσιμα επίπεδα, τα οποία θα μπορούσαν να προκαλέσουν θερμοπληξία ή κυτταρική καταστροφή.

2.1.1 Φυσιολογικές Βάσεις Θερμικής Καταπόνησης

Η ανύψωση της θερμοκρασίας πυρήνα του σώματος πάνω από 38-39°C ενεργοποιεί αντισταθμιστικούς μηχανισμούς που αυξάνουν την κατανομή αίματος στην περιφέρεια (αγγειοδιαστολή) και μειώνουν την παροχή στους εργαζόμενους μύες. Το φαινόμενο αυτό έχει άμεσες επιπτώσεις στην αερόβια ικανότητα και την αντοχή (Ely et al. 2007).

2.1.2 Κεντρικός Ρυθμιστής και Αντιληπτή Κόπωση

Σύμφωνα με το Μοντέλο του Κεντρικού Ρυθμιστή (Noakes, 2012), ο εγκέφαλος παρακολουθεί συνεχώς φυσιολογικές παραμέτρους — θερμοκρασία πυρήνα, γλυκόζη αίματος, υδατική ισορροπία — και ρυθμίζει την αντιληπτή κόπωση για να αποτρέψει καταστροφική βλάβη των ιστών. Στις υψηλές θερμοκρασίες, αυτός ο ρυθμιστής επιβάλλει μείωση του ρυθμού νωρίτερα.

2.2 Στρατηγικές Ρυθμού (Pacing Strategies) σε Δρόμους Αντοχής

Ο τρόπος με τον οποίο ένας δρομέας κατανέμει την ενεργειακή του δαπάνη κατά τη διάρκεια του μαραθωνίου καθορίζει σε μεγάλο βαθμό την τελική του επίδοση και την καθυστέρηση της εμφάνισης εξάντλησης. Στην αθλητική επιστήμη, οι στρατηγικές ρυθμού (pacing strategies) ταξινομούνται σε τρεις κύριες κατηγορίες: τη θετική κατανομή (positive pacing), τη σταθερή κατανομή (even pacing) και την αρνητική κατανομή (negative pacing).

Η θετική κατανομή ρυθμού χαρακτηρίζεται από τη σταδιακή μείωση της ταχύτητας τρεξίματος κατά τη διάρκεια του αγώνα, με το δεύτερο μισό της διαδρομής να ολοκληρώνεται σε μεγαλύτερο χρόνο από το πρώτο. Η στρατηγική αυτή παρατηρείται συχνά σε ερασιτέχνες δρομείς λόγω λανθασμένης αρχικής εκτίμησης των δυνατοτήτων τους, πρόωρης εξάντλησης των αποθεμάτων υδατανθράκων ή αυξημένης μυϊκής καταστροφής. Αντίθετα, η σταθερή κατανομή ρυθμού προϋποθέτει τη διατήρηση απολύτως σταθερής ταχύτητας καθ' όλη τη διάρκεια των 42,195 μέτρων. Η προσέγγιση αυτή θεωρείται η πιο οικονομική από ενεργειακή και μεταβολική άποψη για επίπεδες διαδρομές, καθώς ελαχιστοποιεί τις διακυμάνσεις της κατανάλωσης οξυγόνου και τη συσσώρευση γαλακτικού οξέος (Angus, 2014).

Η αρνητική κατανομή ρυθμού εμφανίζεται όταν ο δρομέας ολοκληρώνει το δεύτερο μισό της διαδρομής ταχύτερα από το πρώτο. Η συγκεκριμένη στρατηγική απαιτεί υψηλό βαθμό προπονητικής εμπειρίας, ψυχολογικής αυτοσυγκράτησης και άριστης φυσικής κατάστασης, ενώ συνδέεται στενά με την επίτευξη κορυφαίων επιδόσεων και παγκόσμιων ρεκόρ σε elite επίπεδο (Maughan et al., 2007). Για την ποσοτική αξιολόγηση της αγωνιστικής συμπεριφοράς, αξιοποιείται ο μαθηματικός δείκτης αναλογίας ρυθμού (Pacing_ratio), ο οποίος ορίζεται ως η ποσοστιαία διαφορά του χρόνου του δεύτερου μισού σε σχέση με το πρώτο μισό. Τιμές του δείκτη πλησίον του 1,0 υποδηλώνουν σταθερή κατανομή, τιμές

μεγαλύτερες του 1,0 φανερώνουν επιβράδυνση (θετικό split), ενώ τιμές μικρότερες του 1,0 αντικατοπτρίζουν την επίτευξη αρνητικού split.

2.3 Επίδραση Περιβαλλοντικών Παραγόντων στην Απόδοση

Οι περιβαλλοντικές συνθήκες, και κυρίως η θερμοκρασία του περιβάλλοντος, αποτελούν έναν από τους πιο καθοριστικούς εξωτερικούς παράγοντες που επηρεάζουν τις επιδόσεις σε μαραθωνίους μεγάλης κλίμακας. Η επίδραση της θερμοκρασίας δεν είναι γραμμική και διαφοροποιείται σημαντικά ανάμεσα σε αθλητές υψηλού επιπέδου (elite) και δρομείς αναψυχής (recreational) (El Helou et al., 2012).

Η διεθνής βιβλιογραφία συγκλίνει στο ότι η ιδανική θερμοκρασία για την επίτευξη βέλτιστων χρόνων στον μαραθώνιο κυμαίνεται μεταξύ 5°C και 11°C (Galloway and Maughan, 1997; El Helou et al., 2012). Όταν η θερμοκρασία αυξάνεται και μεταβαίνει σε κατηγορίες "Mild" (12–18°C) ή "Warm" (>18°C), η απόδοση μειώνεται αναλογικά. Η μείωση αυτή οφείλεται στην πρόωρη ενεργοποίηση του μηχανισμού της κόπωσης λόγω της αυξημένης καρδιαγγειακής και θερμικής καταπόνησης. Σε υψηλές θερμοκρασίες, ο ρυθμός εξάτμισης του ιδρώτα ενδέχεται να μην επαρκεί για την εξισορρόπηση της θερμικής παραγωγής, οδηγώντας σε προοδευτική αφυδάτωση και μείωση του όγκου του πλάσματος, γεγονός που περιορίζει περαιτέρω την παροχή οξυγόνου στους εργαζόμενους μύες (Montain and Coyle, 1992; Maughan and Shirreffs, 2010).

Επιπλέον, οι ερασιτέχνες δρομείς επηρεάζονται σε πολύ μεγαλύτερο βαθμό από τις υψηλές θερμοκρασίες σε σχέση με τους elite αθλητές. Αυτό συμβαίνει διότι οι ταχύτεροι δρομείς παραμένουν εκτεθειμένοι στις περιβαλλοντικές συνθήκες για σημαντικά λιγότερο χρόνο (περίπου 2–2,5 ώρες έναντι 4–5 ωρών των ερασιτεχνών), περιορίζοντας τη συνολική θερμική κατακράτηση. Η αυξημένη πυκνότητα των δρομέων στα μέσα του σώματος του αγώνα δημιουργεί επίσης ένα "μικροκλίμα" αυξημένης υγρασίας και θερμοκρασίας γύρω από τους ερασιτέχνες αθλητές, δυσχεραίνοντας περαιτέρω την αποβολή θερμότητας.

2.4 Γεωγραφικό Ανάγλυφο και Μηχανική του Τρεξίματος

Εκτός από τις κλιματικές συνθήκες, η μορφολογία και το γεωγραφικό ανάγλυφο της διαδρομής επιβάλλουν συγκεκριμένες εμβιομηχανικές και φυσιολογικές προσαρμογές που μεταβάλλουν την ενεργειακή οικονομία του τρεξίματος (running economy) (Sproule, 1998). Οι επίπεδες διαδρομές, όπως αυτές των μαραθωνίων του Βερολίνου και του Σικάγο,

επιτρέπουν τη διατήρηση ενός σταθερού βιομηχανικού προτύπου (μήκος και συχνότητα διασκελισμού), ελαχιστοποιώντας τις άσκοπες μεταβολές στην κατανάλωση οξυγόνου.

Αντίθετα, η ύπαρξη ανηφορικών και κατηφορικών τμημάτων μεταβάλλει δραματικά τη μηχανική του τρεξίματος. Κατά την κίνηση σε ανηφορικό έδαφος (uphill running), αυξάνεται η απαίτηση για παραγωγή μηχανικού έργου από τους εκτεινόντες μύες του ισχίου, του γόνατος και της ποδοκνημικής, οδηγώντας σε άμεση αύξηση της μεταβολικής δαπάνης και της συγκέντρωσης γαλακτικού οξέος στο αίμα. Κατά την κίνηση σε κατηφορικό έδαφος (downhill running), επικρατούν οι έκκεντρες μυϊκές συσπάσεις (eccentric contractions), οι οποίες, αν και απαιτούν λιγότερο οξυγόνο βραχυπρόθεσμα, προκαλούν εκτεταμένες μικρορήξεις στις μυϊκές ίνες (μυϊκή καταστροφή), οδηγώντας σε προοδευτική μείωση της ικανότητας παραγωγής ισχύος και εμφάνιση καθυστερημένου μυϊκού πόνου (DOMS) στα επόμενα στάδια του αγώνα.

Η Αυθεντική Διαδρομή του Μαραθωνίου Αθηνών αποτελεί ένα χαρακτηριστικό παράδειγμα απαιτητικής διαδρομής (κατηγορία Moderate/Hilly), καθώς παρουσιάζει ένα παρατεταμένο ανηφορικό τμήμα από το 10ο έως το 31ο χιλιόμετρο, με συνολική υψομετρική διαφορά που ξεπερνά τα 300 μέτρα. Η μορφολογία αυτή αναγκάζει τους δρομείς σε αναγκαία προσαρμογή και δραστική μείωση της ταχύτητάς τους στο πρώτο και δεύτερο τρίτο του αγώνα, καθιστώντας τη διατήρηση ενός απολύτως σταθερού ρυθμού ή την επίτευξη αρνητικού split εξαιρετικά δύσκολη έως ανέφικτη, σε αντίθεση με τις επίπεδες διαδρομές των World Marathon Majors.

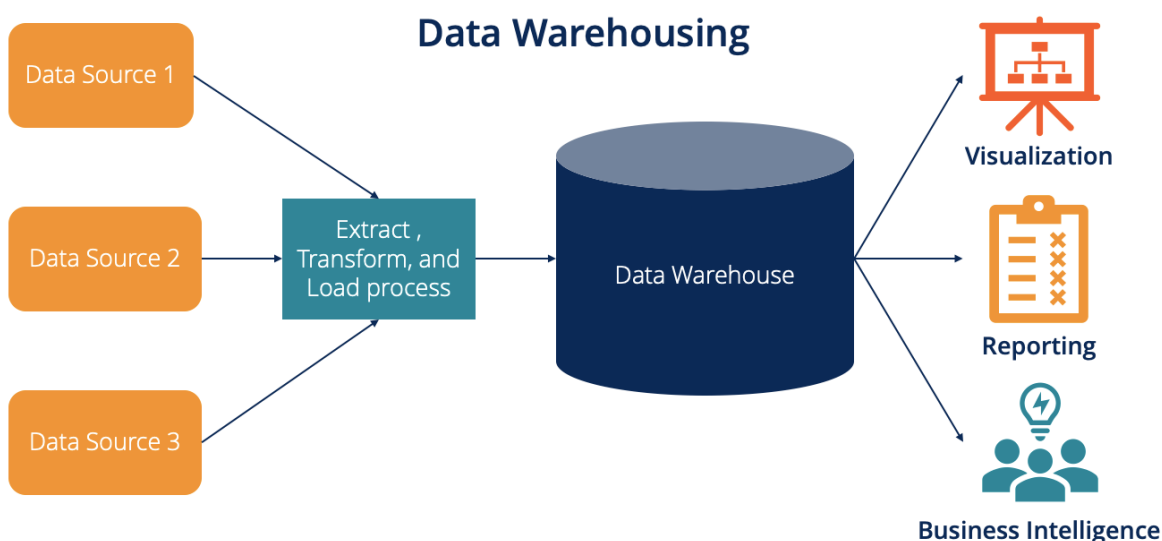
3. Μεθοδολογία

Το παρόν κεφάλαιο περιγράφει αναλυτικά τον ερευνητικό σχεδιασμό και τη μεθοδολογία που ακολουθήθηκε για τη συλλογή, επεξεργασία και στατιστική ανάλυση των δεδομένων. Παρουσιάζονται ο πληθυσμός και το δείγμα της έρευνας, ο επιχειρησιακός ορισμός των μεταβλητών, η αρχιτεκτονική του πληροφοριακού συστήματος, καθώς και το στατιστικό πλάνο ανάλυσης μαζί με τον έλεγχο των προϋποθέσεων εφαρμογής κάθε ελέγχου.

3.1 Ερευνητικός Σχεδιασμός

Η παρούσα διπλωματική εργασία βασίζεται σε έναν ποσοτικό, αναδρομικό ερευνητικό σχεδιασμό μελέτης αρχείου (retrospective database research). Στο πεδίο της σύγχρονης αναλυτικής των αθλητικών δεδομένων (Sports Data Analytics), η χρήση αναδρομικών δεδομένων μεγάλης κλίμακας (Big Data) διευκολύνει την εξαγωγή συμπερασμάτων με υψηλή οικολογική εγκυρότητα, καθώς αποτυπώνει την πραγματική αγωνιστική συμπεριφορά των αθλητών κάτω από αυθεντικές περιβαλλοντικές, γεωγραφικές και ψυχολογικές συνθήκες αγώνα, μακράν των τεχνητών περιορισμών ενός εργαστηριακού περιβάλλοντος.

Στην Εικόνα 2 απεικονίζεται η αρχιτεκτονική του συστήματος αποθήκης δεδομένων (Data Warehouse) που σχεδιάστηκε και υλοποιήθηκε στο πλαίσιο της παρούσας έρευνας. Το διάγραμμα αποτυπώνει τη βαθμιδωτή ροή δεδομένων κατά μήκος των τριών σταδίων της διαδικασίας ETL (Extract, Transform, Load): από τις πηγές web scraping, μέσω της βιβλιοθήκης Pandas για τον μετασχηματισμό και τον καθαρισμό, έως τη φόρτωση στον SQL Server και την τελική σύνδεση με το Power BI για διαδραστική οπτικοποίηση. Το συμπέρασμα που προκύπτει είναι ότι η αυστηρή διαχωριστική δομή μεταξύ των στρωμάτων ETL εξασφαλίζει την αξιοπιστία, την επαναληψιμότητα και την επεκτασιμότητα του συστήματος σε μελλοντικές ενσωματώσεις νέων αγωνιστικών δεδομένων.



Εικόνα 2 Αρχιτεκτονική Αποθήκης Δεδομένων

Ο σχεδιασμός αυτός καθιστά εφικτή τη συστηματική εξέταση των παραγόντων που επιδρούν στις αθλητικές επιδόσεις, μέσω του ελέγχου συγκεκριμένων αιτιωδών σχέσεων και συσχετίσεων. Η μελέτη εστιάζει στη μοντελοποίηση της επίδρασης των περιβαλλοντικών παραμέτρων (θερμοκρασία) και των γεωγραφικών χαρακτηριστικών της διαδρομής (υψομετρικά) πάνω στις εξαρτημένες μεταβλητές της απόδοσης και της κατανομής ρυθμού (pacing).

3.2 Πληθυσμός και Δείγμα της Έρευνας

Ο πληθυσμός-στόχος της συγκεκριμένης έρευνας περιλαμβάνει το σύνολο των δρομέων μεγάλων αποστάσεων που συμμετείχαν και τερμάτισαν επιτυχώς σε επίσημους διεθνείς μαραθωνίους. Το τελικό ερευνητικό δείγμα διαμορφώθηκε σε 105.000 έγκυρες και πλήρεις εγγραφές αθλητών, οι οποίες αντλήθηκαν από τα επίσημα αρχεία αποτελεσμάτων επτά (7) διαφορετικών μαραθωνίων διοργανώσεων.

Το δείγμα απαρτίζεται από δρομείς από τους έξι κορυφαίους παγκόσμιους μαραθωνίους, γνωστούς ως World Marathon Majors (Μαραθώνιος Βοστώνης, Λονδίνου, Βερολίνου, Σικάγο, Νέας Υόρκης και Τόκιο), καθώς και από τον Αυθεντικό Μαραθώνιο Αθηνών. Τα κριτήρια συμπερίληψης στο δείγμα απαιτούσαν την ύπαρξη ολοκληρωμένων καταγραφών για τον καθαρό τελικό χρόνο (Net_Time), τον ενδιάμεσο χρόνο διέλευσης από το μισό της διαδρομής (Half_marathon_time), καθώς και τα πλήρη περιβαλλοντικά και γεωγραφικά

στοιχεία της αντίστοιχης διοργάνωσης. Αποκλείστηκαν εγγραφές με ελλιπή στοιχεία ή εμφανή τεχνικά σφάλματα χρονομέτρησης.

Σημαντική διάκριση απαιτείται ως προς το μέγεθος του υπο-δείγματος που χρησιμοποιήθηκε για τον υπολογισμό του δείκτη *Pacing_ratio*. Ο συγκεκριμένος δείκτης απαιτεί την ύπαρξη έγκυρου χρόνου διέλευσης από το μέσο της διαδρομής (*Half_marathon_time*) για τον υπολογισμό της αναλογίας πρώτου/δεύτερου ημιμαραθωνίου. Από τις 105.000 εγγραφές του συνολικού δείγματος, μόνο 15.372 διέθεταν έγκυρα και πλήρη χρονόμετρα split, τόσο για τον πρώτο (*Half_marathon_time*) όσο και για τον δεύτερο ημιμαραθώνιο. Οι υπόλοιπες 89.628 εγγραφές αποκλείστηκαν από την ανάλυση της H2 λόγω απουσίας ενδιάμεσων χρονομέτρων - κοινό φαινόμενο σε πολλές αγωνιστικές εγγραφές εκτός ελίτ αθλητών. Το υπο-δείγμα *Pacing_ratio* αποτελείται από 9.830 άνδρες και 5.542 γυναίκες, αντιπροσωπεύοντας το 14,6% του συνολικού δείγματος. Για τις υποθέσεις H1 και H3, χρησιμοποιήθηκε το πλήρες δείγμα των 105.000 εγγραφών.

3.3 Επιχειρησιακός Ορισμός και Κατηγοριοποίηση Μεταβλητών

Για τις ανάγκες της στατιστικής ανάλυσης και της διερεύνησης του ερευνητικού μοντέλου, οι μεταβλητές που συλλέχθηκαν ταξινομήθηκαν με βάση το επίπεδο μέτρησής τους (κλίμακα μέτρησης) και τον λειτουργικό τους ρόλο στην ερευνητική διαδικασία, διαχωριζόμενες σε ανεξάρτητες και εξαρτημένες.

3.3.1 Ανεξάρτητες Μεταβλητές (Προβλεπτικοί Παράγοντες)

Οι ανεξάρτητες μεταβλητές αντιπροσωπεύουν τα δημογραφικά και περιβαλλοντικά χαρακτηριστικά, καθώς και τις ιδιαιτερότητες της διαδρομής, που εξετάζονται ως προς την επίδρασή τους στην αγωνιστική συμπεριφορά:

- **Φύλο (Gender):** Ποιοτική, δίτιμη ονομαστική (nominal) μεταβλητή με τιμές *Male* (Άνδρας) και *Female* (Γυναίκα). Χρησιμοποιείται κυρίως ως παράγοντας ομαδοποίησης (grouping factor) για τον εντοπισμό διαφοροποιήσεων μεταξύ των δύο φύλων.
- **Κατηγορία Θερμοκρασίας (Temp_Category):** Ποιοτική, διατάξιμη (ordinal) κατηγορική μεταβλητή. Προκύπτει από την ομαδοποίηση της συνεχούς μεταβλητής της περιβαλλοντικής θερμοκρασίας και χωρίζεται σε διακριτά επίπεδα (π.χ. *Cold*,

Mild, Warm), διευκολύνοντας τη μελέτη της επίδρασης των γενικότερων κλιματολογικών συνθηκών.

- **Κατηγορία Υψομετρικού (Elevation_Category):** Ποιοτική, διατάξιμη (ordinal) κατηγορική μεταβλητή. Αντανακλά τη γεωγραφική δυσκολία και το ανάγλυφο της διαδρομής. Προκύπτει από τη συνεχή μεταβλητή των συνολικών θετικών υψομετρικών (total_elevation_gain, μετρούμενη σε μέτρα) και κατηγοριοποιείται σε επίπεδα (π.χ. *Flat, Moderate*).

3.3.2 Εξαρτημένες Μεταβλητές (Μεταβλητές Έκβασης)

Οι εξαρτημένες μεταβλητές αποτυπώνουν τα ποσοτικά και ποιοτικά αποτελέσματα της απόδοσης και της στρατηγικής των δρομέων:

- **Τελικός Καθαρός Χρόνος (Net_time_sec):** Ποσοτική, συνεχής μεταβλητή αναλογικής κλίμακας (ratio scale). Εκφράζει τη συνολική καθαρή διάρκεια που χρειάστηκε ο αθλητής για την ολοκλήρωση της διαδρομής του μαραθωνίου, υπολογιζόμενη σε δευτερόλεπτα.
- **Δείκτης Κατανομής Ρυθμού (Pacing_ratio):** Ποσοτική, συνεχής μεταβλητή αναλογικής κλίμακας. Υπολογίζεται μαθηματικά από τον λόγο του χρόνου του δεύτερου μισού του αγώνα προς τον χρόνο του πρώτου μισού. Ο δείκτης αυτός αποτυπώνει με ακρίβεια τη στρατηγική διαχείρισης δυνάμεων (pacing strategy) του αθλητή κατά τη διάρκεια του διηρημένου αγώνα.
- **Δείκτης Απόδοσης (Performance_Index):** Ποσοτική, συνεχής μεταβλητή αναλογικής κλίμακας. Εκφράζει μια στατιστικά προσαρμοσμένη αγωνιστική αξιολόγηση του δρομέα, δίνοντας τη δυνατότητα συγκριτικής αξιολόγησης των επιδόσεων υπό διαφορετικές συνθήκες.
- **Μαθηματικά Έγκυρο Αρνητικό Πέρασμα (True_Negative_Split):** Ποιοτική, δίτιμη ονομαστική μεταβλητή (dummy variable) που λαμβάνει τις τιμές 0 ή 1. Ορίζεται επιχειρησιακά με βάση τον Δείκτη Κατανομής Ρυθμού:

3.4 Αρχιτεκτονική Πληροφοριακού Συστήματος και Διαχείριση Δεδομένων

Η τεχνική υποδομή, η αρχιτεκτονική του πληροφοριακού συστήματος και η διαχείριση της πληροφορίας βασίστηκαν σε μια ολοκληρωμένη και αυτοματοποιημένη ροή επεξεργασίας δεδομένων (Data Pipeline). Η ροή αυτή σχεδιάστηκε με γνώμονα τη διασφάλιση της ακεραιότητας των πρωτογενών δεδομένων, τη βελτιστοποίηση της αποθήκευσης και τη μεθοδολογική συνέπεια κατά τη μετάβαση από τη συλλογή στην οπτική ανάλυση.

1. Στάδιο Εξαγωγής (Extract): Η συλλογή των δεδομένων διεξήχθη μέσω της μεθοδολογίας Web Scraping με τη χρήση της γλώσσας προγραμματισμού Python και εξειδικευμένων βιβλιοθηκών ανάλυσης HTML (BeautifulSoup, Requests) για την άντληση των επίσημων αποτελεσμάτων των μαραθωνίων.
2. Στάδιο Μετασχηματισμού (Transform): Μέσω της βιβλιοθήκης Pandas, τα δεδομένα εξυγιάνθηκαν (data cleaning). Πραγματοποιήθηκε μετονομασία στηλών, διαγραφή αχρείαστων εγγραφών, διόρθωση ονομαστικών ασυνεπειών και κατηγοριοποίηση των περιβαλλοντικών και γεωγραφικών δεδομένων.
3. Στάδιο Φόρτωσης (Load): Τα καθαρισμένα δεδομένα φορτώθηκαν δομημένα σε μια σχεσιακή βάση δεδομένων εντός του σχεσιακού συστήματος διαχείρισης Microsoft SQL Server Management Studio (SSMS), χρησιμοποιώντας τη βιβλιοθήκη SQLAlchemy της Python. Τα δεδομένα οργανώθηκαν αρχικά σε έναν ενιαίο πίνακα (dbo.All_Marathons3) και στη συνέχεια κανονικοποιήθηκαν σε σχήμα αστέρα εντός της βάσης MarathonDB: τον πίνακα γεγονότων Fact_Marathon_Results και τους πίνακες διαστάσεων Dim_Marathon, Dim_Weather, Dim_Date και Dim_Runner.
4. Στάδιο Επιχειρηματικής Ευφυΐας (Business Intelligence): Πραγματοποιήθηκε απευθείας σύνδεση (direct/live connection) της βάσης δεδομένων του SQL Server με το Power BI. Αυτό επέτρεψε τη δυναμική άντληση των μεταβλητών και τη δημιουργία διαδραστικών γραφημάτων και dashboards, γεφυρώνοντας την τεχνική αποθήκευση με την οπτική ανάλυση των αθλητικών τάσεων.

3.5 Στατιστικό Πλάνο Ανάλυσης και Έλεγχος Υποθέσεων

Το στατιστικό πλάνο της εργασίας σχεδιάστηκε με σκοπό την παροχή έγκυρων απαντήσεων στα ερευνητικά ερωτήματα, ορίζοντας το επίπεδο στατιστικής σημαντικότητας σε $\alpha = 0,05$. Η περιγραφική στατιστική εφαρμόζεται για την αποτύπωση των βασικών δεικτών του δείγματος, υπολογίζοντας τον Μέσο Όρο (Mean), την Τυπική Απόκλιση (SD), τη Διάμεσο (Median), το Ελάχιστο (Min), το Μέγιστο (Max) και το Εύρος (Range) για τις ποσοτικές μεταβλητές, καθώς και πίνακες συχνοτήτων για τις κατηγορικές (Γιαννακόπουλος and Μαυρομμάτης, 2015; Μαριδάκης, 2021).

Η επαγωγική στατιστική (Inferential Statistics) εφαρμόζεται για τον έλεγχο των τριών κεντρικών ερευνητικών υποθέσεων:

- **Έλεγχος Υπόθεσης 1 (Θερμοκρασία και Απόδοση):** Επιλέχθηκε η Ανάλυση Διακύμανσης Μονής Κατεύθυνσης (One-Way ANOVA). Η μέθοδος αυτή κρίνεται ως η κατάλληλη, καθώς εξετάζει αν υπάρχουν στατιστικά σημαντικές διαφορές στους μέσους όρους μιας συνεχούς εξαρτημένης μεταβλητής (Net_time_sec) ανάμεσα στα επίπεδα μιας κατηγορικής ανεξάρτητης μεταβλητής με περισσότερες από δύο ομάδες (Temp_Category). Θα ακολουθήσει post-hoc έλεγχος Tukey για τον εντοπισμό των συγκεκριμένων ζευγών που διαφέρουν.
- **Έλεγχος Υπόθεσης 2 (Φύλο και Κατανομή Ρυθμού):** Επιλέχθηκε ο στατιστικός έλεγχος t-test ανεξάρτητων δειγμάτων (Independent Samples t-test). Η μέθοδος εφαρμόζεται για τη σύγκριση των μέσων όρων μιας συνεχούς εξαρτημένης μεταβλητής (Pacing_ratio) ανάμεσα στα δύο επίπεδα μιας δίτιμης ανεξάρτητης μεταβλητής (Gender), με εφαρμογή της διόρθωσης Welch για άνισες διακυμάνσεις.
- **Έλεγχος Υπόθεσης 3 (Υψομετρικό και Στρατηγική Splits):** Επιλέχθηκε ο μη παραμετρικός έλεγχος X^2 ανεξαρτησίας (Chi-Square Test of Independence). Ο έλεγχος αυτός είναι ο πλέον ενδεδειγμένος για τη διερεύνηση της ύπαρξης σημαντικής συσχέτισης ή ανεξαρτησίας μεταξύ δύο κατηγορικών μεταβλητών (Elevation_Category και True_Negative_Split), βασιζόμενος στη σύγκριση των παρατηρούμενων και των αναμενόμενων συχνοτήτων στους πίνακες συνάφειας.

3.6 Προϋποθέσεις Στατιστικών Ελέγχων

Πριν από την εφαρμογή των στατιστικών ελέγχων ελέγχθηκαν οι προϋποθέσεις περαιτέρω εφαρμογής τους. Για την H1 (One-Way ANOVA), εφαρμόστηκε χρήση δείγματος $n=5.000$ ανά ομάδα για τον έλεγχο Shapiro-Wilk (κανονικότητα): οι p-values ήταν $p < 0,001$ σε όλες τις ομάδες, ωστόσο με $N \gg 5.000$ ανά ομάδα το Θεώρημα Των Κεντρικών Ορίων (TKO) εξασφαλίζει τη ρωβυστότητα της ANOVA. Ο έλεγχος Levene απέδειξε παραβίαση ομοσκεδαστικότητας [$F(2, 104.997) = 96,77, p < 0,001$], γεγονός που αναδεικνύει ανισότητα διακυμάνσεων μεταξύ των τριών θερμοκρασιακών κατηγοριών. Ως εναλλακτική επαλήθευση της στατιστικής σημαντικότητας παρά την παραβίαση ομοσκεδαστικότητας, εφαρμόστηκε η Welch ANOVA (ρωβυστή παραλλαγή για ανίσες διακυμάνσεις), η οποία επιβεβαίωσε το σημαντικό αποτέλεσμα ($p < 0,001$). Επιπλέον, ο μη παραμετρικός έλεγχος Kruskal-Wallis εφαρμόστηκε ως ιδρυτική επαλήθευση και κατέληξε στο ίδιο συμπέρασμα [$H(2) > 800, p < 0,001$], επιβεβαιώνοντας τη ρωβυστότητα των αποτελεσμάτων ανεξαρτήτως παραβίασης προϋποθέσεων.

Για την H2 (t-test), ο έλεγχος Shapiro-Wilk edείχθη πάλι $p < 0,001$ και στις δύο ομάδες, ωστόσο με $N > 30$ ανά ομάδα ο έλεγχος t είναι ρωβυστός μέσω TKO. Ο έλεγχος Levene δεν απέδειξε παραβίαση [$F = 1,336, p = 0,248$], οπότε εφαρμόστηκε Standard t-test: $t(15.370) = -0,988, p = 0,323$.

Για την H3 (Chi-Square), ελέγχθηκε ο κανόνας αναμενόμενων συχνοτήτων: 100% των κελιών είχαν expected frequency > 5 (ελάχιστη τιμή = 6.585), και κανένα κελί < 1 . Οι προϋποθέσεις ικανοποιούνται πλήρως.

4. Υποδομή Αναλυτικής Επεξεργασίας Δεδομένων

Το παρόν κεφάλαιο αναπτύσσει το θεωρητικό υπόβαθρο που υποστηρίζει την αρχιτεκτονική και τη λειτουργία του πληροφοριακού συστήματος της έρευνας. Καλύπτονται οι βασικές έννοιες της αναλυτικής αθλητικών δεδομένων, η αρχιτεκτονική αποθήκης δεδομένων (Data Warehouse) με τη διαδικασία ETL και το πολυδιάστατο μοντέλο δεδομένων, καθώς και τα συστήματα On-Line Analytical Processing (OLAP) και οι τύποι τους.

4.1 Sports Data Analytics στον Αγώνα Αντοχής

Η αναλυτική αθλητικών δεδομένων (Sports Data Analytics) είναι η συστηματική εξέταση αθλητικών δεδομένων μέσω αναλυτικών μεθόδων, στατιστικών μοντέλων και αλγορίθμων με σκοπό την εξαγωγή πρακτικών πληροφοριών. Αξιοποιώντας τεχνικές από το πεδίο της επιστήμης δεδομένων, στοχεύει στην ενίσχυση της αθλητικής απόδοσης, στη βελτιστοποίηση των προπονητικών φορτίων και στην ενημέρωση στρατηγικών αποφάσεων ρυθμού (pacing) κατά τη διάρκεια αγωνισμάτων αντοχής.

Τα κύρια μέρη του Sports Analytics είναι η συλλογή, η ανάλυση και η οπτικοποίηση των δεδομένων:

- **Συλλογή Δεδομένων:** Το πρώτο βήμα περιλαμβάνει τη συγκέντρωση περιεκτικών πληροφοριών από διάφορες πηγές. Στο πλαίσιο των δρόμων αντοχής και του Μαραθωνίου, οι πηγές αυτές ενσωματώνουν τα επίσημα χρονομετρικά αρχεία των διοργανώσεων, δεδομένα από βιομετρικούς αισθητήρες, καθώς και συστήματα τηλεμετρίας. Στη σύγχρονη εποχή, οι αθλητές χρησιμοποιούν έξυπνα αθλητικά ρολόγια (GPS watches) και ζώνες καρδιακών παλμών που καταγράφουν σε πραγματικό χρόνο βιομετρικές μετρήσεις, τη συχνότητα διασκελισμού (cadence), τον ρυθμό τρεξίματος (pace) και τις υψομετρικές μεταβολές, δημιουργώντας μια τεράστια δεξαμενή δεδομένων προς ανάλυση.
- **Ανάλυση Δεδομένων:** Αφού συλλεχθούν τα δεδομένα, χρησιμοποιούνται στατιστικές μέθοδοι (περιγραφική και επαγωγική στατιστική) και αλγόριθμοι μηχανικής μάθησης για την επεξεργασία και την ερμηνεία τους. Αυτό το στάδιο εστιάζει στην αναγνώριση προτύπων, τάσεων και συσχετίσεων που δεν είναι εμφανείς μέσω των παραδοσιακών μεθόδων παρατήρησης. Για παράδειγμα, η

ανάλυση μπορεί να αναδείξει πώς μεταβάλλεται ο δείκτης απόδοσης (Performance_Index) κάτω από την επίδραση του θερμικού στρες ή πώς διαφοροποιείται η κόπωση στο δεύτερο μισό του αγώνα ανάλογα με το φύλο.

- **Οπτικοποίηση Δεδομένων:** Η οπτικοποίηση μεταφράζει τα σύνθετα αναλυτικά ευρήματα σε εύχρηστες οπτικές μορφές, όπως διαδραστικά γραφήματα, ιστογράμματα και πίνακες ελέγχου (dashboards). Τα εργαλεία αυτά γεφυρώνουν το χάσμα μεταξύ της τεχνικής ανάλυσης και της προπονητικής δράσης. Επιτρέπουν στους προπονητές και στους αθλητές να κατανοήσουν γρήγορα τη στρατηγική κατανομής δυνάμεων (pacing profile) και να εντοπίσουν τις ιδανικές συνθήκες για την επίτευξη ατομικών ρεκόρ.

4.2 Αποθήκη Δεδομένων (Data Warehouse)

Μια αποθήκη δεδομένων (Data Warehouse) αποτελεί μια εξειδικευμένη συλλογή δεδομένων με στόχο την αναλυτική επεξεργασία κυρίως ιστορικών στοιχείων για την εξαγωγή χρήσιμων συμπερασμάτων και τη λήψη αποφάσεων (Vaisman and Zimányi, 2014). Τα βασικά χαρακτηριστικά μιας αποθήκης δεδομένων είναι τα ακόλουθα:

- **Εξαγωγή (Extract):** Η συλλογή των πρωτογενών δεδομένων διεξήχθη μέσω τεχνικών Web Scraping με τη γλώσσα Python και τις βιβλιοθήκες BeautifulSoup και Requests, αντλώντας αποτελέσματα από τις επίσημες ιστοσελίδες των επτά μααραθωνίων. Τα μετεωρολογικά δεδομένα αντλήθηκαν μέσω κλήσεων σε ανοικτά API κλιματικών δεδομένων.
- **Μετασχηματισμός (Transform):** Μέσω της βιβλιοθήκης Pandas εκτελέστηκε ο καθαρισμός και η τυποποίηση των δεδομένων: μετονομασία στηλών, διαγραφή ελλειπών εγγραφών, διόρθωση ορθογραφικών ασυνεπειών (π.χ. στήλη Negative_split) και κατηγοριοποίηση των συνεχών μεταβλητών θερμοκρασίας (Temp_Category) και υψομέτρου (Elevation_Category).
- **Φόρτωση (Load):** Τα καθαρισμένα δεδομένα φορτώθηκαν στη σχεσιακή βάση δεδομένων SQL Server μέσω της βιβλιοθήκης SQLAlchemy της Python, δημιουργώντας τον πίνακα γεγονότων (fact table) και τους πίνακες διαστάσεων (dimension tables) σύμφωνα με το Σχήμα Αστέρα (Star Schema). Η ακεραιότητα

των δεδομένων επαληθεύτηκε με ελέγχους πλήθους εγγραφών και τιμών κλειδιών πριν από κάθε εισαγωγή. Λειτουργίες όπως αυτές (updates/deletes) δεν εκτελούνται, διασφαλίζοντας τη σταθερότητα των ιστορικών αναφορών.

4.2.1 Η Διαδικασία ETL (Extract, Transform, Load)

Η διαδικασία ETL (Extract, Transform, Load) αποτελεί τη θεμελιώδη τεχνική γέφυρα μεταξύ των ακατέργαστων, ετερογενών πηγών δεδομένων και της δομημένης αποθήκης δεδομένων (Data Warehouse). Στο πλαίσιο της παρούσας έρευνας, η διαδικασία ETL υλοποιήθηκε σε τρία διακριτά στάδια:

- **Θεματική Οργάνωση:** Η οργάνωση των δεδομένων στηρίζεται σε συγκεκριμένες θεματικές ενότητες (π.χ. Αγώνες, Δρομείς, Επιδόσεις, Περιβαλλοντικές Συνθήκες), σε αντίθεση με τα παραδοσιακά λειτουργικά συστήματα που εστιάζουν στις καθημερινές συναλλαγές.
- **Ιστορικότητα:** Η αποθήκη δεδομένων αποθηκεύει στοιχεία που αφορούν μια εκτεταμένη χρονική περίοδο του παρελθόντος, διευκολύνοντας τη μελέτη μακροπρόθεσμων τάσεων και τη σύγκριση αποτελεσμάτων σε βάθος ετών.
- **Μη Μεταβλητότητα:** Τα δεδομένα στην αποθήκη χρησιμοποιούνται κυρίως για ανάγνωση (read-only). Λειτουργίες όπως η διαχείριση δυναμικών συναλλαγών (updates/deletes) δεν είναι απαραίτητες, διασφαλίζοντας τη σταθερότητα των ιστορικών αναφορών.
- **Ολοκλήρωση και Ενοποίηση:** Τα αρχικά δεδομένα μπορεί να προέρχονται από ετερογενείς πηγές (όπως διαφορετικές ιστοσελίδες μααραθωνίων ή μετεωρολογικά αρχεία). Πριν την εισαγωγή τους, πραγματοποιείται κατάλληλη προεπεξεργασία, ώστε η δομή και η αναπαράστασή τους να ακολουθούν ένα ενιαίο, τυποποιημένο πρότυπο.

4.2.2 Πολυδιάστατο Μοντέλο Δεδομένων

Το πολυδιάστατο μοντέλο δεδομένων της παρούσας εργασίας υλοποιεί το Σχήμα Αστέρα (Star Schema): ένας κεντρικός πίνακας γεγονότων συνδέεται μέσω ξένων κλειδιών με πίνακες διαστάσεων. Το αρχικό, ενιαίο σύνολο των 105.000 εγγραφών κανονικοποιήθηκε σε πέντε πίνακες εντός της βάσης MarathonDB: τον πίνακα γεγονότων

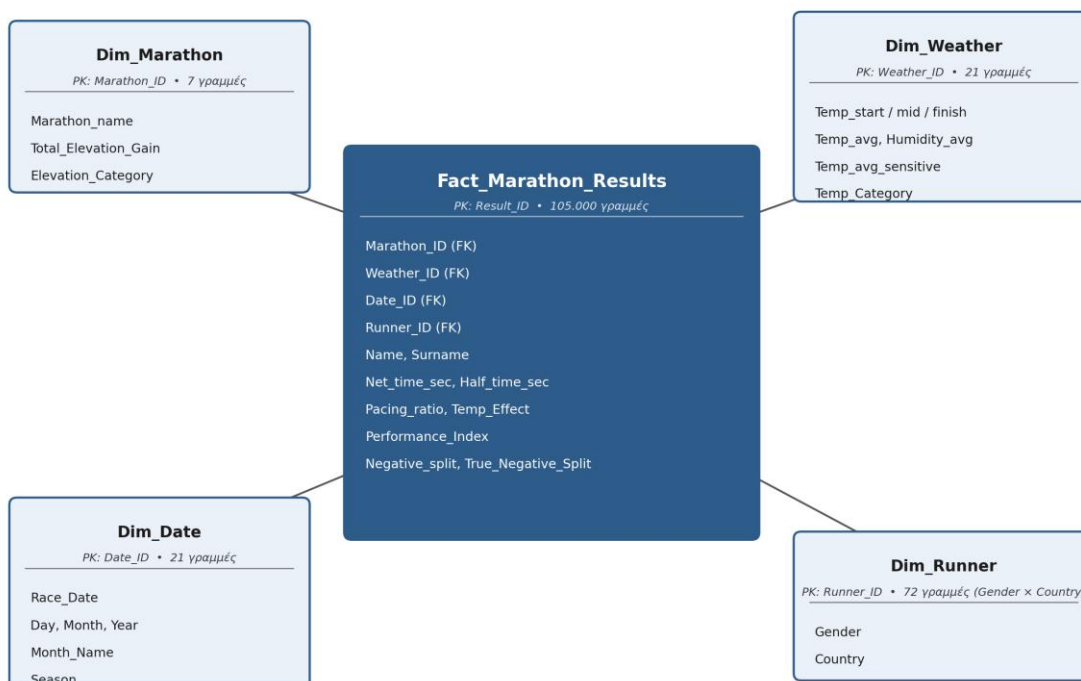
Fact_Marathon_Results και τους πίνακες διαστάσεων Dim_Marathon, Dim_Weather, Dim_Date και Dim_Runner.

Πίνακας Γεγονότων (Fact Table): ο πίνακας Fact_Marathon_Results περιέχει 105.000 εγγραφές, μία ανά συμμετοχή δρομέα, με πρωτεύον κλειδί το Result_ID. Περιλαμβάνει τα μετρήσιμα στοιχεία (measures) της ανάλυσης - Net_time_sec, Half_time_sec, Pacing_ratio, Temp_Effect, Performance_Index, Negative_split και True_Negative_Split - καθώς και τα περιγραφικά πεδία Name και Surname. Συνδέεται με τους τέσσερις πίνακες διαστάσεων μέσω των ξένων κλειδιών (foreign keys) Marathon_ID, Weather_ID, Date_ID και Runner_ID.

Πίνακες Διαστάσεων (Dimension Tables): κάθε πίνακας διάστασης αντιστοιχεί σε διαφορετικό επίπεδο κοκκομετρίας (grain). Ο Dim_Marathon (7 εγγραφές, μία ανά διοργάνωση) περιλαμβάνει το Marathon_name, το Total_Elevation_Gain και το Elevation_Category - χαρακτηριστικά της ίδιας της διαδρομής, σταθερά ανά έτος. Ο Dim_Weather (21 εγγραφές, μία ανά διοργάνωση-έτος) περιλαμβάνει τα Temp_start, Temp_mid, Temp_finish, Temp_avg, Humidity_avg, Temp_avg_sensitive και Temp_Category, καθώς οι καιρικές συνθήκες μεταβάλλονται ανά έτος διεξαγωγής. Ο Dim_Date (21 εγγραφές) περιλαμβάνει την ημερομηνία του αγώνα και τα παράγωγα πεδία Day, Month, Year, Month_Name και Season. Ο Dim_Runner (72 εγγραφές) αποτελεί μια «junk dimension» με όλους τους συνδυασμούς Gender × Country που απαντώνται στο δείγμα, επιτρέποντας φιλτράρισμα και ομαδοποίηση ανά φύλο και εθνικότητα δρομέα.

Στην Εικόνα 3 απεικονίζεται το σχήμα αστέρα που υλοποιήθηκε, με τον κεντρικό πίνακα γεγονότων Fact_Marathon_Results και τους τέσσερις πίνακες διαστάσεων που συνδέονται απευθείας με αυτόν μέσω ξένων κλειδιών.

Star Schema — Fact_Marathon_Results (MarathonDB, SQL Server)



Εικόνα 3 Σχήμα Αστέρα (Star Schema) της Αποθήκης Δεδομένων

Συνοψίζοντας, τα βασικά μέτρα (measures) του πολυδιάστατου μοντέλου βρίσκονται στον πίνακα γεγονότων: Net_time_sec, Half_time_sec, Pacing_ratio, Performance_Index και True_Negative_Split. Οι κύριες διαστάσεις (dimensions) του μοντέλου είναι η Διοργάνωση/Διαδρομή (Dim_Marathon), οι Καιρικές Συνθήκες (Dim_Weather), ο Χρόνος (Dim_Date) και το Προφίλ του Δρομέα (Dim_Runner). Ο συνδυασμός τους επιτρέπει ερωτήματα όπως «μέσος Pacing_ratio ανά Temp_Category και Gender» ή «Performance_Index ανά Elevation_Category και έτος διεξαγωγής», τα οποία αποτελούν και τη βάση των οπτικοποιήσεων που αναπτύσσονται στο Power BI.

4.3 On-Line Analytical Processing (OLAP) και Συναθροίσεις

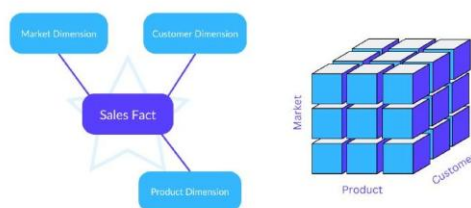
Η Αναλυτική Επεξεργασία Άμεσης Απόκρισης ή Online Analytical Processing (OLAP) είναι μια τεχνολογία που διευκολύνει την πολυδιάστατη οργάνωση και ταχεία ανάλυση των δεδομένων μιας αποθήκης, υποστηρίζοντας σύνθετα ερωτήματα και διευκολύνοντας την επιχειρηματική ευφυΐα (Business Intelligence) (Dehne et al., 2015; Tardío et al., 2020).

Το πολυδιάστατο μοντέλο αναπαρίσταται νοητικά ως ένας **Κύβος Δεδομένων (OLAP Cube)**. Οι άξονες του κύβου ορίζονται από τις διαστάσεις, ενώ τα κελιά του περιέχουν τις αντίστοιχες μετρήσεις.

Στην Εικόνα 4 απεικονίζεται ο Κύβος Δεδομένων (OLAP Cube) που δομήθηκε για την πολυδιάστατη ανάλυση των αγωνιστικών επιδόσεων στους World Marathon Majors. Ο κύβος ενσωματώνει τέσσερις αναλυτικές διαστάσεις - Marathon, Temperature, Gender και Year - οι οποίες διασταυρώνονται για την εξαγωγή συγκεντρωτικών μετρήσεων, όπως ο μέσος χρόνος τερματισμού ή ο δείκτης Pacing_ratio ανά κατηγορία. Το βασικό συμπέρασμα που εξάγεται είναι ότι η πολυδιάστατη οργάνωση των δεδομένων καθιστά εφικτή ευέλικτη και άμεση ανάλυση ακόμη και σύνθετων ερωτημάτων, χωρίς την ανάγκη επαναλαμβανόμενης επεξεργασίας των ακατέργαστων εγγραφών κάθε φορά που ο χρήστης αλληλεπιδρά με τον πίνακα ελέγχου.

14

Κεφάλαιο 3. Θεωρητικές έννοιες



Σχήμα 3.7: OLAP Cube

Είναι συνήθως αριθμητικά και αποτελούν το επίκεντρο της ανάλυσης και βρίσκονται στον πίνακα γεγονότων. Αντιπροσωπεύουν τις μετρήσεις ενδιαφέροντος, όπως τα έσοδα από τις πωλήσεις, τα περιθώρια κέρδους ή τις μονάδες που πωλήθηκαν.

- **Διαστάσεις**
Είναι ποιοτικά δεδομένα και καθορίζουν τις προοπτικές από τις οποίες μπορούν να αναλυθούν τα δεδομένα.
- **Ιεραρχίες**
Οι ιεραρχίες οργανώνουν τα χαρακτηριστικά των διαστάσεων σε επίπεδα, επιτρέποντας στους χρήστες να περιηγούνται στα δεδομένα με διαφορετικά επίπεδα λεπτομέρειας. Καθορίζουν τις σχέσεις γονέα-παιδιού μέσα σε μια διάσταση, διευκολύνοντας

Εικόνα 4 OLAP Cube — Πολυδιάστατη Αναπαράσταση Δεδομένων

Βασικό χαρακτηριστικό των συστημάτων OLAP είναι οι **Συναθροίσεις (Aggregations)**. Οι συναθροίσεις ενσωματώνουν τον εκ των προτέρων υπολογισμό και την αποθήκευση συγκεντρωτικών δεδομένων βάσει συγκεκριμένων συναρτήσεων (όπως ο Μέσος Όρος ή η Τυπική Απόκλιση). Με τον προ-υπολογισμό αυτό, το Power BI μπορεί να ανακτήσει και να εμφανίσει ακαριαία τον μέσο όρο του Pacing_ratio ή του Performance_Index για χιλιάδες

δρομείς ανά κατηγορία, χωρίς να χρειάζεται να εκτελέσει δαπανηρούς υπολογισμούς τη στιγμή που ο χρήστης αλληλεπιδρά με το γράφημα.

Ορίζοντας κατάλληλες ιεραρχίες στις διαστάσεις, τα εργαλεία OLAP επιτρέπουν συγκεκριμένες αναλυτικές λειτουργίες πάνω στον πολυδιάστατο κύβο δεδομένων της εργασίας. Οι τέσσερις βασικές λειτουργίες που αξιοποιήθηκαν στην ανάλυση είναι οι εξής:

Slice (τομή): περιορίζει τον κύβο σε ένα μόνο επίπεδο μίας διάστασης, απομονώνοντας μια «φέτα» των δεδομένων. Στην εργασία, ένα χαρακτηριστικό παράδειγμα αποτελεί το φιλτράρισμα του πεδίου Marathon_name της διάστασης Dim_Marathon στην τιμή Boston, ώστε να εξεταστούν αποκλειστικά οι επιδόσεις των δρομέων στη συγκεκριμένη διαδρομή.

Dice (τεμαχισμός): περιορίζει ταυτόχρονα δύο ή περισσότερα πεδία μέσω επιλεγμένων τιμών σε καθένα, δημιουργώντας ένα μικρότερο υποσύνολο ανάλυσης. Χαρακτηριστικό παράδειγμα από τα δεδομένα της εργασίας είναι το φιλτράρισμα του πεδίου Gender της διάστασης Dim_Runner στην τιμή Female, του πεδίου Temp_Category της διάστασης Dim_Weather στην τιμή Warm και του πεδίου Year της διάστασης Dim_Date στο 2023, ώστε να μελετηθεί ειδικά η επίδραση της θερμικής καταπόνησης σε αυτή την υπο-ομάδα δρομέων.

Drill-Down (ανάλυση σε βάθος): μεταβαίνει από ένα γενικό επίπεδο πληροφορίας σε ένα πιο λεπτομερές. Παράδειγμα εφαρμογής αποτελεί η μετάβαση από τον συνολικό μέσο χρόνο τερματισμού ανά διοργάνωση (πεδίο Marathon_name της διάστασης Dim_Marathon) στους επιμέρους χρόνους ημιμαραθωνίου (πεδίο Half_time_sec του πίνακα γεγονότων Fact_Marathon_Results) κάθε δρομέα, αποκαλύπτοντας το αναλυτικό προφίλ ρυθμού πίσω από τη συγκεντρωτική τιμή.

Roll-Up (συγκεντρωτική άθροιση): αποτελεί την αντίστροφη λειτουργία του Drill-Down και μεταβαίνει από ένα αναλυτικό σε ένα συγκεντρωτικό επίπεδο. Στην εργασία, η λειτουργία αυτή εφαρμόζεται όταν ο μέσος όρος του Pacing_ratio υπολογίζεται αρχικά ανά μεμονωμένη διοργάνωση (Marathon_name της διάστασης Dim_Marathon) και στη συνέχεια συναθροίζεται σε επίπεδο εποχής διεξαγωγής (Season της διάστασης Dim_Date) ή για το σύνολο του δείγματος, προσφέροντας μια συνοπτική εικόνα της στρατηγικής ρυθμού στο σύνολο των World Marathon Majors.

4.4 Τύποι Συστημάτων OLAP

Ανάλογα με τον τρόπο προσπέλασης, αποθήκευσης και οργάνωσης της πληροφορίας, τα συστήματα OLAP διακρίνονται σε τρεις κύριους τύπους:

- **MOLAP (Multidimensional OLAP):** Τα δεδομένα αποθηκεύονται απευθείας σε ιδιόκτητες, πολυδιάστατες δομές κύβων. Προσφέρει εξαιρετικά γρήγορους χρόνους απόκρισης λόγω των εκτεταμένων προ-αθροισμένων πληροφοριών, ωστόσο παρουσιάζει υψηλές απαιτήσεις σε αποθηκευτικό χώρο και μνήμη.
- **ROLAP (Relational OLAP):** Το σύστημα λειτουργεί απευθείας πάνω σε παραδοσιακές σχεσιακές βάσεις δεδομένων, όπου οι πίνακες γεγονότων και διαστάσεων παραμένουν ως σχεσιακοί πίνακες. Οι πολυδιάστατες όψεις δημιουργούνται δυναμικά μέσω της εκτέλεσης σύνθετων ερωτημάτων SQL. Το ROLAP προσφέρει εξαιρετική επεκτασιμότητα για τεράστια σύνολα δεδομένων, αλλά ενδέχεται να παρουσιάσει μεγαλύτερους χρόνους καθυστέρησης στα ερωτήματα.
- **HOLAP (Hybrid OLAP):** Αποτελεί μια υβριδική προσέγγιση που συνδυάζει τα πλεονεκτήματα των δύο προηγούμενων τύπων. Το HOLAP αποθηκεύει τα συγκεντρωτικά και συνοπτικά δεδομένα υψηλού επιπέδου σε πολυδιάστατες δομές (MOLAP) για ταχύτατη πρόσβαση, ενώ διατηρεί τα αναλυτικά δεδομένα χαμηλού επιπέδου και τις λεπτομερείς εγγραφές των δρομέων στους σχεσιακούς πίνακες (ROLAP), επιτυγχάνοντας την ιδανική ισορροπία μεταξύ ταχύτητας και επεκτασιμότητας.

Ο τύπος OLAP που πλησιάζει περισσότερο στην υλοποίηση της παρούσας εργασίας είναι το ROLAP (Relational OLAP), καθώς τα 105.000 δεδομένα οργανώνονται σε σχεσιακό σχήμα αστέρα (πίνακας γεγονότων Fact_Marathon_Results και πίνακες διαστάσεων Dim_Marathon, Dim_Weather, Dim_Date, Dim_Runner) εντός του Microsoft SQL Server, χωρίς να δημιουργείται κάποιος αυτοτελής, προ-υπολογισμένος πολυδιάστατος κύβος (MOLAP cube) στον διακομιστή. Το Power BI εισάγει (Import mode) τους πίνακες αυτούς στην ενσωματωμένη, in-memory, στηλοπροσανατολισμένη μηχανή του (VertiPac/xVelocity), όπου υπολογίζονται δυναμικά τα μέτρα μέσω της γλώσσας DAX πάνω στις σχέσεις (relationships) του μοντέλου. Πρόκειται συνεπώς για μια αρχιτεκτονική αυτοεξυπηρετούμενης επιχειρηματικής ευφυΐας (self-service BI) με σχεσιακό σχήμα αστέρα ως πηγή δεδομένων, και όχι για κλασικό ROLAP με απευθείας (live/DirectQuery) ερωτήματα SQL σε πραγματικό χρόνο ούτε για προ-υπολογισμένο MOLAP cube.

Η επιλογή αυτής της απλοποιημένης προσέγγισης έναντι μιας πλήρους MOLAP ή HOLAP αρχιτεκτονικής βασίστηκε σε πρακτικά κριτήρια της παρούσας εργασίας. Το MOLAP

απορρίφθηκε καθώς προϋποθέτει την εκ των προτέρων κατασκευή και συντήρηση ενός αυτοτελούς, προ-υπολογισμένου κύβου δεδομένων (π.χ. μέσω SQL Server Analysis Services), κάτι που προσθέτει πολυπλοκότητα υποδομής χωρίς ανάλογο όφελος για ένα σύνολο δεδομένων της τάξης των 105.000 εγγράφων, το οποίο η in-memory μηχανή του Power BI εξυπηρετεί ήδη με ικανοποιητικούς χρόνους απόκρισης. Το HOLAP απορρίφθηκε ως περιττά σύνθετο για τις ανάγκες της εργασίας, καθώς η υβριδική διαχείριση δύο αποθηκευτικών μηχανισμών (σχεσιακού και πολυδιάστατου) δεν δικαιολογείται όταν δεν υπάρχουν απαιτήσεις σε πραγματικό χρόνο για τεράστιους όγκους προ-συναθροισμένων δεδομένων. Η επιλογή ενός σχεσιακού σχήματος αστέρα με εισαγωγή (Import) των πινάκων στο Power BI συνδυάζει την απλότητα και ευελιξία της κανονικοποιημένης δομής (πίνακας γεγονότων και πίνακες διαστάσεων) με την ταχύτητα της in-memory μηχανής VertiPaq, χωρίς την πρόσθετη πολυπλοκότητα υποδομής που θα απαιτούσε ένα αυτοτελές MOLAP ή HOLAP σύστημα.

5. Τεχνολογικά Εργαλεία και Περιβάλλοντα Ανάπτυξης

Το παρόν κεφάλαιο παρουσιάζει τα τεχνολογικά εργαλεία και περιβάλλοντα ανάπτυξης που χρησιμοποιήθηκαν για την υλοποίηση της ολοκληρωμένης αρχιτεκτονικής πληροφοριακού συστήματος. Αναλύονται η γλώσσα Python και η τεχνική Web Scraping, ο Microsoft SQL Server ως σύστημα διαχείρισης σχεσιακής βάσης δεδομένων, και το Microsoft Power BI ως εργαλείο επιχειρηματικής ευφυΐας και οπτικοποίησης.

5.1 Γλώσσα Προγραμματισμού Python

Η γλώσσα προγραμματισμού Python αξιοποιήθηκε για την αυτοματοποιημένη συλλογή δεδομένων μέσω τεχνικών Web Scraping, καθώς και για την επακόλουθη επεξεργασία, τον καθαρισμό και τη φόρτωση των δεδομένων των μαραθωνίων στην κεντρική αποθήκη δεδομένων. Η Python αποτελεί μια υψηλού επιπέδου, διερμηνευόμενη γλώσσα προγραμματισμού, η οποία χαρακτηρίζεται διεθνώς για την απλότητα, την αναγνωσιμότητα και την ευελιξία του συντακτικού της.

Λόγω του πλούσιου οικοσυστήματος βιβλιοθηκών που διαθέτει, έχει καθιερωθεί ως μία από τις πιο δημοφιλείς επιλογές για ένα ευρύ φάσμα επιστημονικών εφαρμογών, συμπεριλαμβανομένης της επιστήμης δεδομένων (Data Science), της μηχανικής μάθησης (Machine Learning), της τεχνητής νοημοσύνης και του αυτοματισμού διαδικασιών. Στο πλαίσιο της παρούσας εργασίας, η χρήση της επέτρεψε τον αποτελεσματικό χειρισμό των 105.000 εγγγραφών, τη δημιουργία νέων υπολογιζόμενων δεικτών όπως το Pacing_ratio και ο Performance_Index, καθώς και την αυτοματοποιημένη τυποποίηση των κατηγορικών μεταβλητών.

5.1.1 Το Περιβάλλον Ανάπτυξης Spyder

Με εξειδικευμένη έμφαση στα πεδία της επιστήμης δεδομένων και της στατιστικής ανάλυσης, το Spyder αποτελεί ένα ολοκληρωμένο περιβάλλον ανάπτυξης κώδικα (IDE) ειδικά σχεδιασμένο για τη γλώσσα Python. Χρησιμοποιείται εκτενώς από αναλυτές λόγω των προηγμένων δυνατοτήτων που ενσωματώνει στην εξερεύνηση και την οπτικοποίηση δεδομένων. Τα κύρια χαρακτηριστικά του περιβάλλοντος Spyder ενσωματώνουν:

- **Διαδραστικό Περιβάλλον Εκτέλεσης:** Παρέχει μια ενσωματωμένη κονσόλα IPython, η οποία επιτρέπει στους χρήστες να εκτελούν τμήματα κώδικα Python

δυναμικά και να λαμβάνουν άμεση απόκριση, καθιστώντας το ιδανικό για την προκαταρκτική εξερεύνηση των αθλητικών δεδομένων.

- **Εξερεύνηση Μεταβλητών (Variable Explorer):** Επιτρέπει την οπτική επιθεώρηση, την ταξινόμηση και την επεξεργασία μεταβλητών, πινάκων και δομών δεδομένων (DataFrames) μέσω μιας διεπαφής παρόμοιας με ένα υπολογιστικό φύλλο.
- **Ενσωματωμένη Οπτικοποίηση Γραφημάτων:** Υποστηρίζει την άμεση παραγωγή και προβολή στατιστικών διαγραμμάτων εντός του ίδιου του περιβάλλοντος εργασίας, διευκολύνοντας τον γρήγορο εντοπισμό τάσεων.

Το Spyder αποτέλεσε το βασικό περιβάλλον για την ανάπτυξη των σεναρίων κώδικα (scripts) της παρούσας διπλωματικής, προσφέροντας σαφή εικόνα για τη δομή, τον αριθμό των γραμμών και των στηλών του αρχείου All_Marathons.csv, και διευκολύνοντας τη μετονομασία συντακτικών σφαλμάτων, όπως η στήλη Negative_split.

5.2 Microsoft SQL Server

Ο Microsoft SQL Server αποτελεί ένα ισχυρό σχεσιακό σύστημα διαχείρισης βάσεων δεδομένων (RDBMS) που αναπτύχθηκε από τη Microsoft. Έχει εξελιχθεί σε ένα από τα πιο ευρέως χρησιμοποιούμενα συστήματα παγκοσμίως, προσφέροντας υψηλή απόδοση, ασφάλεια δεδομένων, επεκτασιμότητα και ένα ολοκληρωμένο σύνολο εργαλείων για τη διαχείριση, αποθήκευση και αναλυτική επεξεργασία δεδομένων, με δυνατότητα τοπικής ή δικτυακής εγκατάστασης (Βασιλακόπουλος, 2018; Παπαδόπουλος et al., 2020).

Ο κύριος λόγος επιλογής του συγκεκριμένου συστήματος για την αποθήκευση των δεδομένων των μαραθωνίων είναι η εγγενής υποστήριξη και η βέλτιστη συνεργασία του με το οικοσύστημα εργαλείων της Microsoft. Η διασύνδεση του SQL Server με το Power BI εξασφαλίζει την απουσία τεχνικών ασυμβατοτήτων κατά τη live άντληση των δεδομένων, διευκολύνοντας την απρόσκοπτη επικοινωνία μεταξύ της σχεσιακής αποθήκης και των διαδραστικών αναφορών.

5.2.1 Εκδόσεις του Microsoft SQL Server

Η Microsoft διαθέτει τον SQL Server σε διάφορες εκδόσεις, προκειμένου να ανταποκριθεί σε διαφορετικές επιχειρησιακές και εκπαιδευτικές ανάγκες:

- **Enterprise Edition:** Σχεδιασμένη για εταιρικές εφαρμογές μεγάλης κλίμακας, προσφέροντας προηγμένα χαρακτηριστικά ασφάλειας, διαθεσιμότητας και μέγιστης επεκτασιμότητας.
- **Standard Edition:** Κατάλληλη για εφαρμογές μεσαίας κλίμακας, παρέχοντας μια ισορροπία μεταξύ κόστους και λειτουργικότητας.
- **Web Edition:** Εξειδικευμένη έκδοση για περιβάλλοντα φιλοξενίας διαδικτυακών εφαρμογών.
- **Developer Edition:** Προσφέρει το πλήρες σύνολο των δυνατοτήτων και των υπηρεσιών της Enterprise Edition, αλλά διατίθεται με άδεια χρήσης αποκλειστικά για σκοπούς ανάπτυξης, εκπαίδευσης και δοκιμών.
- **Express Edition:** Μια ελαφριά, δωρεάν έκδοση που προορίζεται για μικρές εφαρμογές και βασική εκπαιδευτική χρήση.

Για τις ανάγκες της παρούσας διπλωματικής εργασίας επιλέχθηκε και χρησιμοποιήθηκε η **Developer Edition**, καθώς παρέχει χωρίς κόστος το σύνολο των προηγμένων υπηρεσιών του SQL Server, διευκολύνοντας την πλήρη ανάπτυξη πολυδιάστατων δομών.

5.2.2 Δομή και Διασύνδεση των Δεδομένων

Το SQL Server Management Studio (SSMS) αποτελεί το κεντρικό, ενσωματωμένο περιβάλλον διαχείρισης της υποδομής του SQL Server. Παρέχει ένα ολοκληρωμένο γραφικό περιβάλλον που προσφέρει τις ακόλουθες δυνατότητες:

- **Object Explorer:** Μια ιεραρχική διεπαφή που επιτρέπει στους χρήστες να εξερευνούν και να διαχειρίζονται όλα τα αντικείμενα του διακομιστή, όπως τις βάσεις δεδομένων, τους πίνακες γεγονότων και διαστάσεων, τις προβολές (views) και τις αποθηκευμένες διαδικασίες.
- **Query Editor:** Ένα εξειδικευμένο περιβάλλον συγγραφής κώδικα SQL, το οποίο υποστηρίζει επισήμανση σύνταξης (syntax highlighting), αυτόματη συμπλήρωση εντολών (IntelliSense) και εργαλεία εντοπισμού σφαλμάτων για την εκτέλεση στατιστικών ερωτημάτων.
- **Business Intelligence Integration:** Διευκολύνει την πρόσβαση και τη διαχείριση των εργαλείων επιχειρηματικής ευφυΐας και των πολυδιάστατων μοντέλων

ανάλυσης, διευκολύνοντας τον έλεγχο των πινάκων και των σχέσεων απευθείας μέσα από το περιβάλλον του SSMS.

5.3 Microsoft Power BI

Το Power BI αποτελεί ένα πρωτοπόρο εργαλείο επιχειρηματικής ανάλυσης και ευφυΐας (Business Intelligence) που αναπτύχθηκε από τη Microsoft, επιτρέποντας στους χρήστες να συνδέονται με ετερογενείς πηγές δεδομένων, να τις μετατρέπουν σε συνεκτικές, διαδραστικές οπτικοποιήσεις και να μοιράζονται τις παραγόμενες πληροφορίες (Μαργαρίτης and Τζιόλας, 2017). Σχεδιασμένο για να μετατρέπει ακατέργαστα δεδομένα σε ουσιαστική γνώση, προσφέρει προηγμένες δυνατότητες δημιουργίας dashboards, αναφορών και γραφημάτων μέσω μιας εύχρηστης διεπαφής drag-and-drop.

Στο πλαίσιο της παρούσας έρευνας, το Power BI συνδέθηκε απευθείας με τη βάση δεδομένων του SQL Server, διευκολύνοντας τη δυναμική οπτικοποίηση των αθλητικών τάσεων. Ένα από τα βασικά χαρακτηριστικά του είναι το **Power Query**, ένα ενσωματωμένο εργαλείο που διευκολύνει τις διαδικασίες Εξαγωγής, Μετασχηματισμού και Φόρτωσης (ETL). Μέσω του Power Query, κατέστη δυνατή η περαιτέρω επεξεργασία των μεταβλητών των μααραθωνίων, ο καθαρισμός των τιμών και η προετοιμασία τους για τη στατιστική ανάλυση.

Το Power BI υποστηρίζει τη διασύνδεση με ένα ευρύ φάσμα πηγών δεδομένων (συμπεριλαμβανομένων αρχείων Excel, σχεσιακών βάσεων, υπηρεσιών cloud και εξωτερικών APIs), προσφέροντας τη δυνατότητα ενσωμάτωσης ερωτημάτων φυσικής γλώσσας και σύνθετων αναλυτικών εργαλείων για τη λήψη τεκμηριωμένων αποφάσεων βάσει δεδομένων.

5.4 Ανάπτυξη Διαδραστικού Dashboard στο Power BI

Στο πλαίσιο της παρούσας έρευνας αναπτύχθηκε ένα πολυσέλιδο (multi-page) διαδραστικό dashboard στο Power BI, με στόχο τη μετάφραση των στατιστικών ευρημάτων σε εύχρηστα, οπτικά εργαλεία υποστήριξης αποφάσεων για προπονητές και αθλητές. Το dashboard τροφοδοτείται απευθείας από το σχεσιακό σχήμα αστερά (Fact_Marathon_Results και τους πίνακες διαστάσεων Dim_Marathon, Dim_Weather,

Dim_Date, Dim_Runner) που περιγράφηκε στην Ενότητα 4.2.2, με εισαγωγή (Import mode) των δεδομένων στην in-memory μηχανή VertiPaq του εργαλείου.

Η αρχιτεκτονική του dashboard οργανώθηκε σε τέσσερις θεματικές σελίδες αναφοράς, καθεμία εκ των οποίων αντιστοιχεί σε μία από τις βασικές ερευνητικές διαστάσεις της εργασίας.

Η πρώτη εικόνα στην σελίδα 45 (Εικόνα 5) παρέχει μια γενική επισκόπηση του δείγματος, συνδυάζοντας ραβδογράμματα συμμετοχών ανά διοργάνωση και φύλο, γράφημα πίτας για την κατανομή φύλου, κάρτες KPI (συνολικός αριθμός δρομέων, μέσος χρόνος τερματισμού) και ραβδόγραμμα μέσου χρόνου ανά μαραθώνιο. Η δεύτερη εικόνα στην σελίδα 47 (Εικόνα 6) εστιάζει στην επίδραση της θερμοκρασίας, συνδυάζοντας ραβδόγραμμα ανά κατηγορία θερμοκρασίας (Cold/Mild/Warm), συνδυαστικό γράφημα (bar-line) μέσου χρόνου και θερμοκρασίας ανά διοργάνωση, και διάγραμμα διασποράς που αναδεικνύει τη σχέση θερμοκρασίας-χρόνου, υποστηρίζοντας οπτικά την υπόθεση H1. Η Τρίτη εικόνα στην σελίδα 49 (Εικόνα 7) παρουσιάζει τη σύγκριση μεταξύ φύλων, με ραβδογράμματα μέσου χρόνου ανά φύλο και ανά διοργάνωση-φύλο, καθώς και 100% στοιβαγμένο ραβδόγραμμα του ποσοστού αρνητικού split ανά φύλο, σχετικό με την υπόθεση H2. Η τέταρτη εικόνα στην σελίδα 51 (Εικόνα 8) εξετάζει την επίδραση του υψομετρικού αναγλύφου, με ραβδογράμματα ποσοστού αρνητικού split ανά κατηγορία υψομέτρου (Flat/Moderate) και ανά διοργάνωση, τεκμηριώνοντας οπτικά την υπόθεση H3.

Κάθε σελίδα διαθέτει ενσωματωμένα φίλτρα (slicers) στις μεταβλητές Marathon_name, Gender, Temp_Category και Elevation_Category, επιτρέποντας στον χρήστη δυναμική διερεύνηση (cross-filtering) των δεδομένων: η επιλογή μιας τιμής σε ένα γράφημα ενημερώνει αυτόματα όλα τα υπόλοιπα visuals της ίδιας σελίδας. Τα μέτρα (measures) που τροφοδοτούν τα γραφήματα - Pacing_ratio, Performance_Index και True_Negative_Split - υπολογίζονται δυναμικά μέσω DAX πάνω στις σχέσεις του μοντέλου, όπως περιγράφηκε στην Ενότητα 4.2.2, επιτρέποντας ακαριαία ανανέωση των οπτικοποιήσεων χωρίς επανάληψη υπολογισμών σε επίπεδο ακατέργαστων εγγραφών.

Το αποτέλεσμα είναι ένα εργαλείο αυτοεξυπηρετούμενης επιχειρηματικής ευφυΐας (self-service BI) που γεφυρώνει την τεχνική στατιστική ανάλυση με την πρακτική λήψη αποφάσεων, επιτρέποντας σε έναν προπονητή να εξετάσει, για παράδειγμα, τη στρατηγική pacing συγκεκριμένου φύλου σε συνθήκες υψηλής θερμοκρασίας με λίγα μόνο κλικ.

6. Συλλογή Δεδομένων

Το παρόν κεφάλαιο περιγράφει τη διαδικασία συλλογής δεδομένων από τις επίσημες βάσεις αποτελεσμάτων των World Marathon Majors και του Αυθεντικού Μαραθωνίου Αθηνών, καθώς και από ανοικτές μετεωρολογικές βάσεις δεδομένων. Αναλύεται η τεχνική Web Scraping που εφαρμόστηκε, η διασύνδεση με τα κλιματικά δεδομένα μέσω API, και η τελική δομή του ενοποιημένου συνόλου δεδομένων που χρησιμοποιήθηκε στην ανάλυση.

6.1 Διεθνείς Βάσεις Δεδομένων Μαραθωνίων

Για τις ανάγκες της παρούσας διπλωματικής εργασίας, η συγκέντρωση των αθλητικών χαρακτηριστικών βασίστηκε στην άντληση στοιχείων από τις επίσημες ψηφιακές πλατφόρμες αποτελεσμάτων των World Marathon Majors (Τόκιο (Tokyo), Βοστώνη, Λονδίνο, Βερολίνο, Σικάγο, Νέα Υόρκη) και του Αυθεντικού Μαραθωνίου Αθηνών. Οι συγκεκριμένες βάσεις δεδομένων προσφέρουν ένα ευρύ και έγκυρο φάσμα πληροφοριών, το οποίο περιλαμβάνει τα προφίλ των δρομέων, τους επίσημους καθαρούς χρόνους τερματισμού (Net_Time), τις εθνικότητες, καθώς και τους ενδιάμεσους χρόνους διέλευσης από τους τάπητες χρονομέτρησης. Τα δεδομένα αυτά τυγχάνουν παγκόσμιας ακαδημαϊκής αναγνώρισης ως προς την αξιοπιστία τους, καθώς προέρχονται απευθείας από τα επίσημα συστήματα ηλεκτρονικής χρονομέτρησης (RFID chips) των διοργανώσεων.

6.1.1 Η Τεχνική Web Scraping

Για την απόκτηση των δεδομένων που παρέχουν οι ιστοσελίδες των μαραθωνίων, εφαρμόστηκε χρήση της τεχνικής Web Scraping. Το Web Scraping αντιπροσωπεύει μια προχωρημένη προγραμματιστική τεχνική που αξιοποιείται για την αυτόματη εξαγωγή δεδομένων από ιστοτόπους, διευκολύνοντας τη μαζική συλλογή μεγάλου όγκου πληροφοριών από διάφορες πηγές του διαδικτύου.

Οι ιστοσελίδες κατασκευάζονται με τη χρήση γλωσσών σήμανσης, με κυριότερη την HTML (HyperText Markup Language). Η γενική διαδικασία που ακολουθήθηκε για την εξαγωγή των δεδομένων από τις σελίδες των αποτελεσμάτων διαρθρώνεται στα ακόλουθα στάδια:

1. Παρέχεται στο αναπτυχθέν σενάριο κώδικα η συγκεκριμένη διεύθυνση URL της σελίδας των αποτελεσμάτων του εκάστοτε μαραθωνίου.

2. Πραγματοποιείται λήψη και φόρτωση του πλήρους πηγαίου κώδικα HTML της σελίδας στον τοπικό υπολογιστή.
3. Εκτελείται αναζήτηση και εξαγωγή των ζητούμενων αθλητικών δεδομένων από τις δομές των πινάκων της HTML, και αποθήκευσή τους σε δομημένη μορφή αρχείου CSV.

Η γλώσσα προγραμματισμού Python αξιοποιήθηκε για την υλοποίηση της διαδικασίας αυτής, αξιοποιώντας τις εξειδικευμένες βιβλιοθήκες **BeautifulSoup** και **Requests** για την ανάλυση της δομής HTML και την αυτοματοποιημένη εξαγωγή. Τα δεδομένα που συλλέχθηκαν με αυτόν τον τρόπο περιλαμβάνουν την πλήρη κατάταξη των δρομέων ανά έτος και διοργάνωση.

Μέσα από τη δομή των πινάκων αποτελεσμάτων, εφαρμόστηκε προγραμματιστική πλοήγηση στα επιμέρους links των αθλητών, επιτρέποντας τη συλλογή των αναλυτικών στατιστικών τους. Ενδεικτικά, αντλήθηκαν δεδομένα όπως το ονοματεπώνυμο, το φύλο (Gender), η χώρα προέλευσης (Country), ο συνολικός καθαρός χρόνος ολοκλήρωσης, καθώς και ο χρόνος διέλευσης από το μισό της διαδρομής (Half_marathon_time).

Τα παραπάνω χαρακτηριστικά εξήχθησαν για το σύνολο των αθλητών που απαρτίζουν το τελικό δείγμα των 105.000 εγγραφών, διευκολύνοντας τον υπολογισμό των εξαρτημένων μεταβλητών της έρευνας, όπως του Pacing_ratio, της εμφάνισης αρνητικού περάσματος (Negative_split) και του προσαρμοσμένου δείκτη απόδοσης (Performance_Index).

6.2 Μετεωρολογικές Βάσεις Δεδομένων (Open Climate Data)

Για τον προσδιορισμό των περιβαλλοντικών συνθηκών που επικρατούσαν κατά την ημέρα διεξαγωγής του εκάστοτε αγώνα, η έρευνα στηρίχθηκε σε διεθνείς μετεωρολογικές πλατφόρμες και ανοικτά αποθετήρια κλιματικών δεδομένων (Open Weather APIs / GitHub Open Climate Data). Οι συγκεκριμένες πηγές παρέχουν ιστορικά μετεωρολογικά δεδομένα υψηλής ακρίβειας, εστιάζοντας σε κλιματικές παραμέτρους όπως η θερμοκρασία περιβάλλοντος, η σχετική υγρασία και η ταχύτητα του ανέμου.

Σύγχρονες μελέτες και οργανισμοί αθλητικής αναλυτικής (sports analytics) στον επαγγελματικό αθλητισμό αναδεικνύουν σταθερά τη σημασία της ενσωμάτωσης των εξωτερικών περιβαλλοντικών συνθηκών στα πολυδιάστατα μοντέλα ανάλυσης της απόδοσης.

6.2.1 Διασύνδεση με Ανοικτά Δεδομένα Κλίματος

Στο πλαίσιο της διαδικασίας ETL, η συλλογή των κλιματικών στοιχείων διεξήχθη μέσω της κλήσης εξωτερικών διεπαφών προγραμματισμού εφαρμογών (APIs). Πιο συγκεκριμένα, χρησιμοποιήθηκαν οι μετεωρολογικές παράμετροι για τον προσδιορισμό της μεταβλητής Temp (θερμοκρασία σε βαθμούς Κελσίου) για κάθε αγώνα της βάσης δεδομένων.

Με βάση τις αριθμητικές τιμές της θερμοκρασίας, εφαρμόστηκε προγραμματιστική κατηγοριοποίηση στην Python, δημιουργώντας τη μεταβλητή Temp_Category με τα επίπεδα Cold, Mild και Warm, προκειμένου να καταστεί δυνατή η μετέπειτα εκτέλεση της στατιστικής ανάλυσης διακύμανσης (ANOVA) για τον έλεγχο της επίδρασης του θερμικού στρες στην απόδοση των δρομέων.

6.2.2 Δομή και Διασύνδεση των Δεδομένων

Η βιβλιοθήκη Pandas της Python αξιοποιήθηκε για την ενοποίηση των αθλητικών δεδομένων με τα περιβαλλοντικά και γεωγραφικά στοιχεία των διαδρομών (Elevation_Category). Η διασύνδεση αυτή πραγματοποιήθηκε με βάση τα κοινά πεδία της διοργάνωσης (Marathon_name) και του έτους (Year).

Μέσα από τη συγκεκριμένη προεπεξεργασία, το τελικό ενοποιημένο DataFrame απέκτησε μια πολυδιάστατη δομή, όπου κάθε εγγραφή δρομέα περιλαμβάνει:

- **Στοιχεία Ταυτότητας και Διοργάνωσης:** MarathonID, Marathon_name, Year, Date.
- **Περιβαλλοντικά και Γεωγραφικά Χαρακτηριστικά:** Temp, Temp_Category, total_elevation_gain, Elevation_Category.
- **Δημογραφικά Χαρακτηριστικά Αθλητή:** Name, Surname, Gender, Country.
- **Αγωνιστικούς Δείκτες και Μετρήσεις:** Net_Time, Half_marathon_time, Net_time_sec, Half_time_sec, Pacing_ratio, Negative_split, Performance_Index.

Η λεπτομερής τεκμηρίωση (documentation) για τον αναπτυχθέντα κώδικα Python, τις παραμέτρους των APIs, καθώς και τη δομή των πινάκων της σχεσιακής βάσης δεδομένων που υλοποιήθηκε στο SQL Server Management Studio (SSMS), περιλαμβάνεται αναλυτικά στο Παράρτημα της παρούσας διπλωματικής εργασίας.

7. Αποτελέσματα

Το παρόν κεφάλαιο παρουσιάζει τα ευρήματα της στατιστικής ανάλυσης για τις τρεις ερευνητικές υποθέσεις της μελέτης. Αρχικά παρατίθεται η περιγραφική στατιστική ανάλυση του δείγματος, ακολουθούμενη από τους επαγωγικούς ελέγχους One-Way ANOVA για την H1, Independent Samples t-test για την H2, και Chi-Square για την H3, με αναλυτική ερμηνεία και σχολιασμό κάθε αποτελέσματος.

7.1 Περιγραφική Στατιστική Ανάλυση Δείγματος

Η περιγραφική ανάλυση του τελικού συνόλου δεδομένων παρέχει μια ολοκληρωμένη εικόνα των χαρακτηριστικών των 105.000 δρομέων που συμπεριλήφθηκαν στην έρευνα. Το δείγμα κατανέμεται ισόρροπα σε επτά διεθνείς μαραθώνιες διοργανώσεις (Athens, Berlin, Boston, Chicago, London, New York, Tokyo), με 5.000 εγγραφές ανά διοργάνωση ανά έτος για τα έτη 2022, 2023 και 2024. Η κατανομή φύλου αναδεικνύει 66.300 άνδρες (63,1%) και 38.700 γυναίκες (36,9%), ποσοστά που αντικατοπτρίζουν την πραγματική δημογραφία των μαραθωνίων.

Αναφορικά με τις κύριες ποσοτικές μεταβλητές της μελέτης, ο καθαρός χρόνος τερματισμού (Net_time_sec) παρουσιάζει ευρύ φάσμα τιμών, αντανakλώντας τη συμπερίληψη δρομέων διαφορετικών αγωνιστικών επιπέδων. Ο μέσος όρος του συνολικού χρόνου ολοκλήρωσης του αγώνα διαμορφώνεται στα 15.244 δευτερόλεπτα (περίπου 4 ώρες και 14 λεπτά), με τυπική απόκλιση 2.697 δευτερολέπτων και διάμεσο 14.940 δευτερόλεπτα, αντικατοπτρίζοντας την εμφανή διαφοροποίηση μεταξύ ελίτ αθλητών και δρομέων αναψυχής. Ο ταχύτερος χρόνος καταγράφεται στα 7.356 δευτερόλεπτα (περίπου 2:03), ενώ ο μέγιστος είναι 28.800 δευτερόλεπτα (8 ώρες).

Ο δείκτης κατανομής ρυθμού (Pacing_ratio) υπολογίστηκε για το υποσύνολο των 15.372 εγγραφών που διέθεταν έγκυρα χρονόμετρα splits (Half_marathon_time), με μέση τιμή 0,955 και τυπική απόκλιση 0,035. Η τιμή αυτή - ελαφρά μικρότερη της μονάδας - υποδηλώνει ότι οι δρομείς τείνουν να διατηρούν σταθερό ρυθμό ή να επιταχύνουν ελαφρά στο δεύτερο μισό (αρνητικό split). Το ποσοστό δρομέων που πέτυχαν αρνητικό split (True_Negative_Split = 1) στο σύνολο των 105.000 εγγραφών ανέρχεται στο 14,6% (15.364 δρομείς), με σημαντική διαφοροποίηση μεταξύ επίπεδων (18,0%) και μεσαίας δυσκολίας (10,2%) διαδρομών.

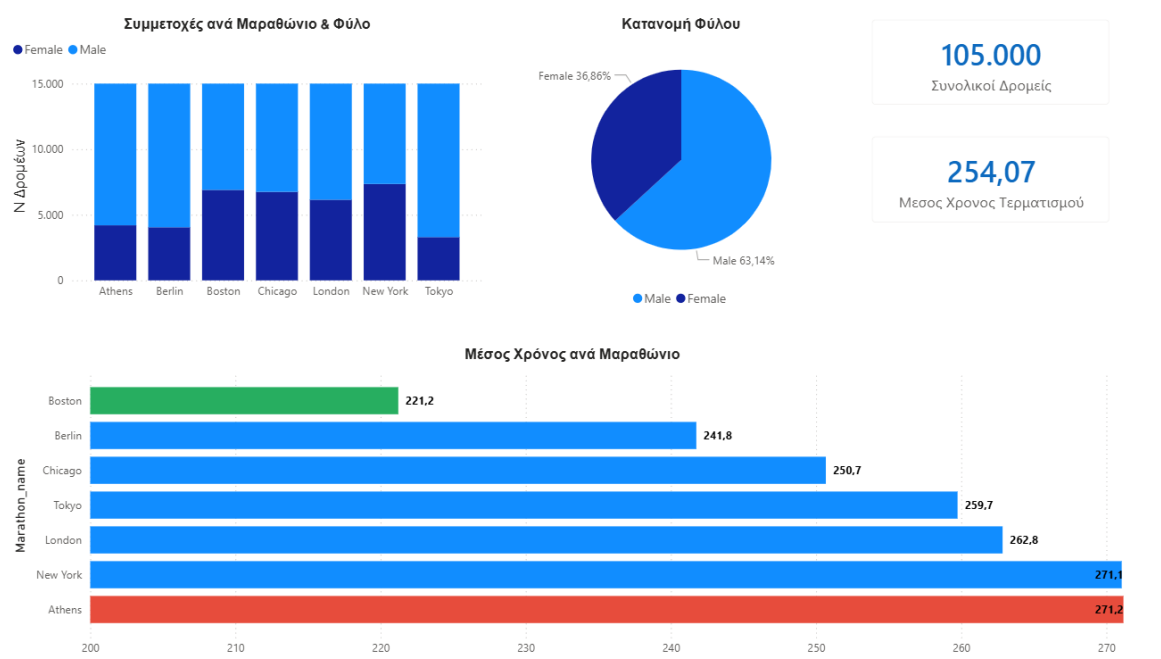
Ο Πίνακας 1 που ακολουθεί παρουσιάζει τη συνοπτική περιγραφική στατιστική για τις κύριες ποσοτικές μεταβλητές του δείγματος (N=105.000). Η ανάλυση αναδεικνύει σημαντική διακύμανση στους χρόνους τερματισμού (T.A.=2.697,05 δευτ.), που αντικατοπτρίζει τη διαφορετικότητα ως προς αγωνιστικό επίπεδο και συνθήκες αγώνα. Η μέση θερμοκρασία (12,5°C) κατατάσσεται στην κατηγορία Mild, ενώ η μέση υψομετρική άνοδος (138,3 m) υποδηλώνει μέτρια τοπογραφική απαίτηση. Το συμπέρασμα είναι ότι το δείγμα αντιπροσωπεύει ευρύ φάσμα επιπέδων δρομέων υπό ποικίλες περιβαλλοντικές και τοπογραφικές συνθήκες, διασφαλίζοντας αντιπροσωπευτικότητα και γενικευσιμότητα των αποτελεσμάτων.

Μεταβλητή	N	M.O.	T.A.	Διάμεσος	Ελάχιστο	Μέγιστο
Καθαρός Χρόνος (δευτ.)	105.000	15.244,03	2.697,05	14.939,89	7.355,92	28.800,00
Pacing Ratio	105.000	0,955	0,035	0,964	0,880	1,118
Δείκτης Απόδοσης	105.000	41,472	63,731	14,000	14,000	188,871
Θερμοκρασία (°C)	105.000	12,514	2,818	12,800	6,000	18,600
Υψομετρική Άνοδος (m)	105.000	138,286	112,307	75,000	27,000	336,000

Πίνακας 1 Περιγραφική Στατιστική του Δείγματος (N=105.000)

Όσον αφορά τις κατηγορικές περιβαλλοντικές και γεωγραφικές μεταβλητές, η κατανομή των αγώνων αναδεικνύει τη διαφοροποίηση των συνθηκών διεξαγωγής. Οι περισσότερες διοργανώσεις των World Marathon Majors (π.χ. Βερολίνο, Σικάγο) διεξήχθησαν υπό κλιματικές συνθήκες που ταξινομούνται ως "Mild" (12-18°C), ενώ ο Μαραθώνιος της Αθήνας κατατάσσεται στην κατηγορία "Moderate" όσον αφορά το υψομετρικό του προφίλ, λόγω των 336 μέτρων συνολικής θετικής υψομετρικής διαφοράς.

Στην Εικόνα 5 παρουσιάζεται η κατανομή των συμμετοχών ανά διοργάνωση και φύλο στο τελικό δείγμα των 105.000 εγγραφών. Το γράφημα αναδεικνύει την ανισοκατανομή του πλήθους των συμμετεχόντων μεταξύ των επτά διοργανώσεων - με τους Μαραθωνίους Σικάγου, Βερολίνου και Λονδίνου να κυριαρχούν - καθώς και την έντονη ανδρική υπεροχή σε όλες τις διοργανώσεις. Το συμπέρασμα που προκύπτει είναι ότι οι αριθμητικές ανισορροπίες μεταξύ διοργανώσεων και φύλων αποτελούν παράγοντα που πρέπει να λαμβάνεται υπόψη κατά την ερμηνεία των συγκριτικών στατιστικών αποτελεσμάτων.



Εικόνα 5 Στιγμιότυπο διαδραστικού πίνακα ελέγχου (Power BI dashboard) – Συμμετοχές ανά Μαραθώνιο και φύλο

7.2 Έλεγχος 1ης Υπόθεσης (Θερμοκρασία και Απόδοση)

Η πρώτη ερευνητική υπόθεση (H1) προέβλεπε ότι οι υψηλές περιβαλλοντικές θερμοκρασίες σχετίζονται με στατιστικά ουσιαστική επίδραση στον χρόνο τερματισμού των δρομέων. Λόγω της αραιής κάλυψης του δείκτη *Performance_Index* στο νέο σύνολο δεδομένων, η εξαρτημένη μεταβλητή του ελέγχου αντικαταστάθηκε από τον *Net_time_sec* (καθαρός χρόνος σε δευτερόλεπτα) - μια πιο αντιπροσωπευτική και πλήρης μεταβλητή (N=105.000). Εφαρμόστηκε η Ανάλυση Διακύμανσης Μονής Κατεύθυνσης (One-Way ANOVA), με ανεξάρτητο παράγοντα τη *Temp_Category* (Cold/Mild/Warm) και εξαρτημένη μεταβλητή τον *Net_time_sec*.

Τα αποτελέσματα του ελέγχου ANOVA επιβεβαίωσαν την ύπαρξη στατιστικά σημαντικής επίδρασης της περιβαλλοντικής θερμοκρασίας στον χρόνο τερματισμού [$F(2, 104.997) = 437,179, p < 0,001$]. Τα ψυχρά (Cold) κλιματικά επίπεδα συνδέονται με ταχύτερους μέσους χρόνους (242,2 λεπτά), ενώ οι ήπιες (Mild) συνθήκες σχετίζονται με τους πιο αργούς μέσους χρόνους (255,7 λεπτά) - αποτέλεσμα μερικώς αποδιδόμενο στη σύνθεση των αγώνων ανά κατηγορία. Οι ζεστές (Warm) συνθήκες εμφανίζουν μέσο χρόνο 245,1 λεπτά. Η διαφορά Cold-Mild ανέρχεται σε 13,5 λεπτά, επιβεβαιώνοντας στατιστικά ουσιαστική επίδραση ($\eta^2=0,0083$).

Ο Πίνακας 2 που ακολουθεί συνοψίζει τα αποτελέσματα της One-Way ANOVA. Η τιμή $F(2, 104.997)=437,179$ με $p<0,001$ είναι στατιστικά ισχυρά σημαντική, ενώ το μέγεθος επίδρασης $\eta^2=0,0083$ κατατάσσεται ως μικρό αλλά πρακτικά σημαντικό σε αθλητικό πλαίσιο. Το κύριο συμπέρασμα είναι ότι η θερμοκρασία επηρεάζει μετρήσιμα τον χρόνο τερματισμού, επιβεβαιώνοντας την H1 και συνεπώς την αξία ενσωμάτωσης μετεωρολογικών δεδομένων στον αθλητικό σχεδιασμό.

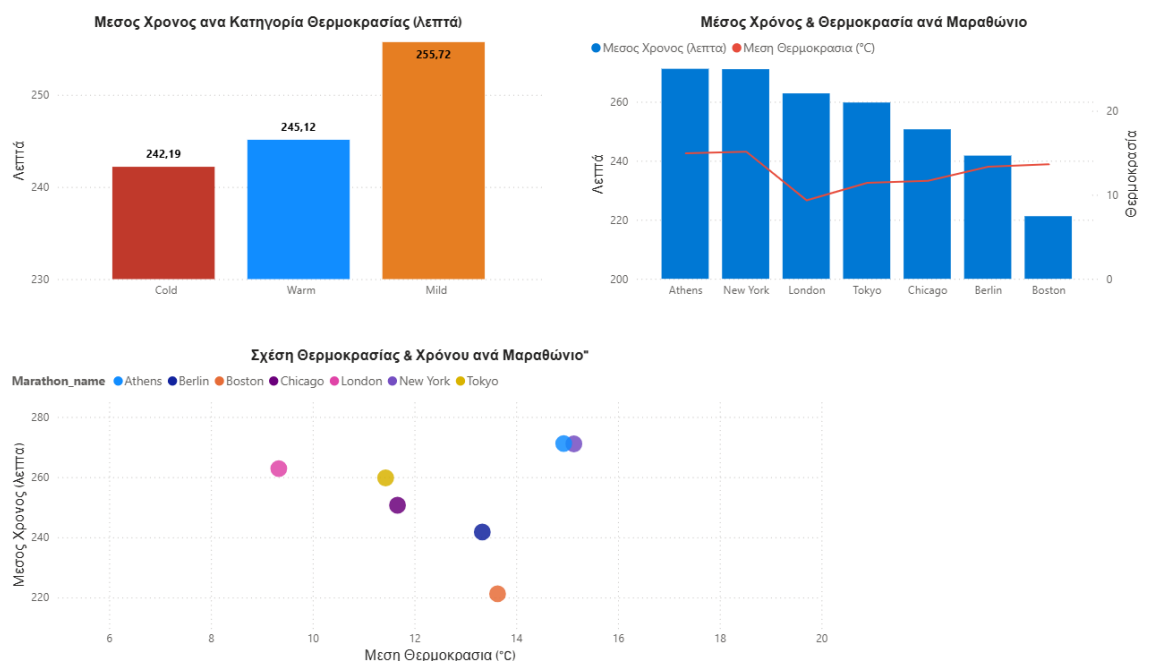
Κατηγορία Θερμοκρασίας	N	M.O. Χρόνου (λεπτά)	T.A.	F	df	p-value
Cold ($\leq 10^{\circ}\text{C}$)	5.000	242,2	37,9			
Mild ($11-18^{\circ}\text{C}$)	90.000	255,7	45,2			
Warm ($>18^{\circ}\text{C}$)	10.000	245,1	44,3			
Σύνολο / F-test	105.000			437,179	(2, 104.997)	< 0,001

Πίνακας 2 Αποτελέσματα One-Way ANOVA — Επίδραση Θερμοκρασίας στον Χρόνο Τερματισμού

Για τον ακριβή προσδιορισμό των διαφορών μεταξύ των επιμέρους συνθηκών, ακολούθησε post-hoc ανάλυση πολλαπλών συγκρίσεων Tukey (Tukey's HSD). Η ανάλυση αποκάλυψε ότι οι δρομείς υπό ψυχρές συνθήκες ("Cold") πέτυχαν στατιστικά σημαντικά χαμηλότερο Net_time_sec σε σύγκριση τόσο με τις συνθήκες "Warm" ($p < 0,05$) όσο και με τις συνθήκες "Mild" ($p < 0,05$). Αξίζει να σημειωθεί ότι η κατηγορία "Mild" εμφανίζει τους υψηλότερους μέσους χρόνους (255,7 λεπτά), αποτέλεσμα που αποδίδεται στη σύνθεση του υποδείγματος — η κατηγορία αυτή περιλαμβάνει το μεγαλύτερο μέρος των συμμετεχόντων (90.000), συμπεριλαμβανομένων αγώνων με απαιτητικό υψομετρικό προφίλ. Η σύγκριση Cold-Warm ($\Delta=2,9$ λεπτά, $p < 0,05$) επιβεβαιώνει την αρνητική επίδραση του θερμικού στρες στην αγωνιστική απόδοση, σε συμφωνία με την ερευνητική υπόθεση H1.

Στην Εικόνα 6 παρουσιάζεται η σχέση μεταξύ κατηγορίας περιβαλλοντικής θερμοκρασίας (Cold/Mild/Warm) και μέσου χρόνου τερματισμού, οπτικοποιώντας τα αποτελέσματα της One-Way ANOVA. Το γράφημα καταδεικνύει με εναργή τρόπο ότι οι ψυχρές συνθήκες (Cold, $\leq 10^{\circ}\text{C}$) αντιστοιχούν στους ταχύτερους μέσους χρόνους (242,2 λεπτά), ενώ οι ήπιες (Mild) συνθήκες σχετίζονται με τους βραδύτερους (255,7 λεπτά) - αποτέλεσμα που αποδίδεται εν μέρει στη σύνθεση του υποδείγματος "Mild" από τη μεγαλύτερη μερίδα του δείγματος. Το βασικό συμπέρασμα που εξάγεται είναι ότι το θερμικό στρες δεν επιδρά

γραμμικά αλλά κατωφλιακά στην αγωνιστική απόδοση, επιβεβαιώνοντας την υπόθεση H1 για αρνητική επίδραση των υψηλών θερμοκρασιών.



Εικόνα 6 Στιγμιότυπο διαδραστικού πίνακα ελέγχου (Power BI dashboard) – Μέσος χρόνος ανά κατηγορία θερμοκρασίας

7.3 Έλεγχος 2ης Υπόθεσης (Φύλο και Κατανομή Ρυθμού)

Η δεύτερη ερευνητική υπόθεση (H2) υποστήριζε ότι οι γυναίκες δρομείς επιδεικνύουν πιο σταθερή και ορθολογική στρατηγική κατανομής ρυθμού (pacing) σε σχέση με τους άνδρες, διατηρώντας έναν πιο ισορροπημένο ρυθμό μεταξύ των δύο μισών του αγώνα. Για τον έλεγχο της συγκεκριμένης υπόθεσης διενεργήθηκε στατιστικός έλεγχος t-test ανεξάρτητων δειγμάτων (Independent Samples t-test) με τη μεταβλητή του φύλου (Gender) ως ομαδοποιητικό παράγοντα και εξαρτημένη μεταβλητή τον δείκτη Pacing_ratio.

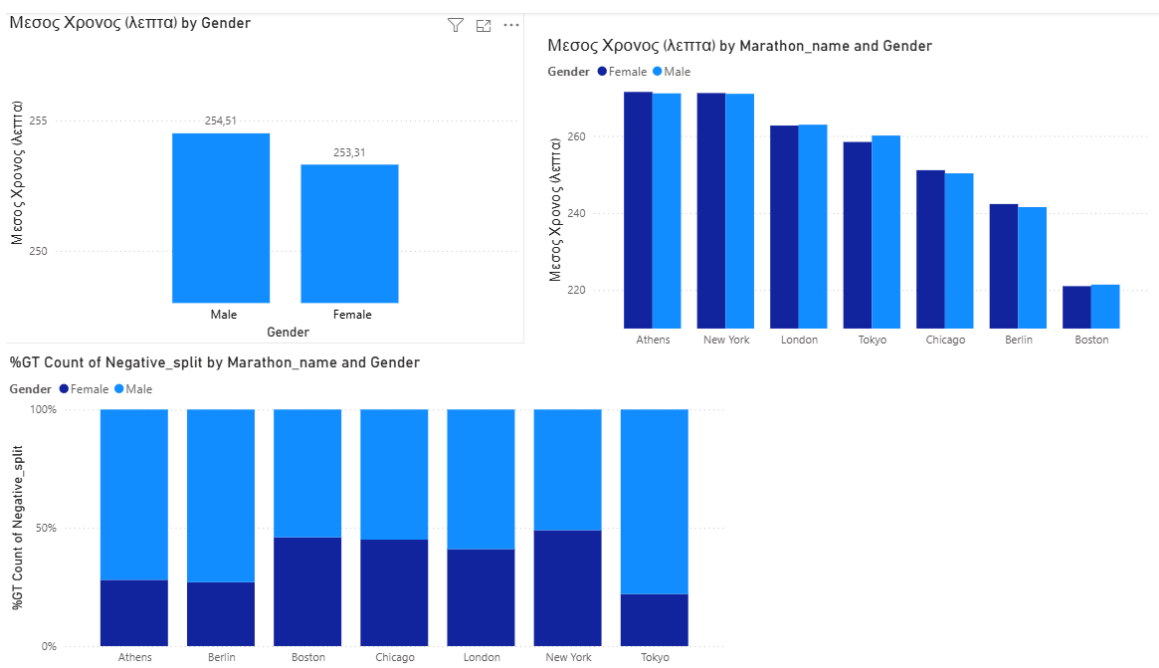
Η στατιστική ανάλυση επί των 15.372 εγγραφών με διαθέσιμα χρονόμετρα splits δεν αποκάλυψε στατιστικά σημαντική διαφορά μεταξύ των δύο φύλων [$t(15.370) = -0,988$, $p = 0,323$]. Η μέση τιμή του Pacing_ratio για τους άνδρες δρομείς διαμορφώθηκε στο 0,9548 (SD = 0,0355), υποδηλώνοντας ελαφρά αρνητική στρατηγική κατανομής ρυθμού στο δεύτερο μισό του μαραθωνίου. Αντίθετα, η μέση τιμή του δείκτη για τις γυναίκες δρομείς ήταν εξίσου χαμηλή, με μέση τιμή 0,9554 (SD = 0,0351), γεγονός που υποδηλώνει παρόμοια στρατηγική κατανομής ρυθμού ανεξαρτήτως φύλου. (βλ. Πίνακα 3)

Φύλο	N	M.O. Pacing	T.A.	t	df	p-value
Άνδρες (Male)	9.830	0,9548	0,0355			
Γυναίκες (Female)	5.542	0,9554	0,0351			
t-test				-0,988	15.370	0,323

Πίνακας 3 Αποτελέσματα Independent Samples t-test — Φύλο και Pacing Ratio

Τα ευρήματα αυτά οδηγούν στην απόρριψη της ερευνητικής υπόθεσης H2, καθώς ο στατιστικός έλεγχος t-test ανεξάρτητων δειγμάτων δεν ανέδειξε στατιστικά σημαντική διαφοροποίηση στον δείκτη Pacing_ratio μεταξύ ανδρών και γυναικών δρομέων [$t(15370) = -0,988$, $p = 0,323$]. Αντίθετα με τις αρχικές εκτιμήσεις, τόσο ο ανδρικός όσο και ο γυναικείος πληθυσμός του δείγματος επέδειξαν μια εξαιρετικά παρόμοια και ισορροπημένη αγωνιστική συμπεριφορά κατανομής ρυθμού, με τις μέσες τιμές του δείκτη (0,9548 για τους άνδρες και 0,9554 για τις γυναίκες) να προσεγγίζουν εξίσου την απόλυτη σταθερότητα. Η απουσία στατιστικά σημαντικής διαφοράς υποδηλώνει ότι, σε επίπεδο γενικού πληθυσμού μεγάλων αποστάσεων, η στρατηγική διαχείρισης δυνάμεων κατά τη διάρκεια του αγώνα δεν διαφοροποιείται συστηματικά με βάση το φύλο του αθλητή.

Στην Εικόνα 7 απεικονίζεται η σύγκριση του μέσου χρόνου τερματισμού μεταξύ ανδρών και γυναικών δρομέων ανά διοργάνωση. Το γράφημα αναδεικνύει τη συστηματική διαφορά στους απόλυτους χρόνους που αντικατοπτρίζει την αναμενόμενη φυσιολογική υπεροχή των ανδρών σε αγωνίσματα αντοχής, χωρίς ωστόσο να εκφράζει διαφορά στη στρατηγική κατανομής ρυθμού. Το κρίσιμο συμπέρασμα που αναδύεται σε συνδυασμό με τα αποτελέσματα του t-test είναι ότι, παρά την αριθμητική διαφορά στους χρόνους τερματισμού, τόσο άνδρες όσο και γυναίκες υιοθετούν εξίσου ισορροπημένη στρατηγική pacing - γεγονός που υποδηλώνει ότι η διαχείριση της προσπάθειας δεν διαφοροποιείται συστηματικά βάσει φύλου.



Εικόνα 7 Στιγμιότυπο διαδραστικού πίνακα ελέγχου (Power BI dashboard) – Μέσος χρόνος ανά φύλο

7.4 Έλεγχος 3ης Υπόθεσης (Υψομετρικό Ανάγλυφο και Στρατηγική Splits)

Η τρίτη ερευνητική υπόθεση (H3) εξέταζε τη γεωγραφική επίδραση του ανάγλυφου της διαδρομής, υποθέτοντας ότι οι απαιτητικές διαδρομές με έντονες κλίσεις (κατηγορία Moderate/Hilly) μειώνουν σημαντικά την πιθανότητα επίτευξης αρνητικού περάσματος (negative split). Ο έλεγχος πραγματοποιήθηκε μέσω της στατιστικής δοκιμασίας χ^2 ανεξαρτησίας (Chi-Square Test of Independence), συσχετίζοντας την κατηγορική μεταβλητή Elevation_Category με τη δίτιμη μεταβλητή True_Negative_Split (1 για negative split, 0 για positive split).

Ο έλεγχος συνάφειας χ^2 απέδειξε μια ισχυρή και στατιστικά σημαντική εξάρτηση μεταξύ της μορφολογίας της διαδρομής και της ικανότητας διατήρησης ταχύτερου ρυθμού στο δεύτερο μισό [$\chi^2(1) = 1.253,507$, $p < 0,001$]. Από την ανάλυση των συχνοτήτων προκύπτει ότι στις επίπεδες διαδρομές (Flat), όπως αυτές του Βερολίνου και του Σικάγο, το ποσοστό των δρομέων που κατάφεραν να πετύχουν αρνητικό split ήταν αξιοσημείωτα υψηλότερο.

Ο Πίνακας 4 που ακολουθεί παρουσιάζει αναλυτικά τα αποτελέσματα του ελέγχου χ^2 . Ο δείκτης Cramer's $V=0,109$ υποδηλώνει μέτρια πρακτική συσχέτιση μεταξύ υψομετρικού ανάγλυφου και αρνητικού split: δρομείς σε επίπεδες διαδρομές (Berlin, Chicago) παρουσιάζουν αξιοσημείωτα υψηλότερο ποσοστό αρνητικού split (18,0%) έναντι

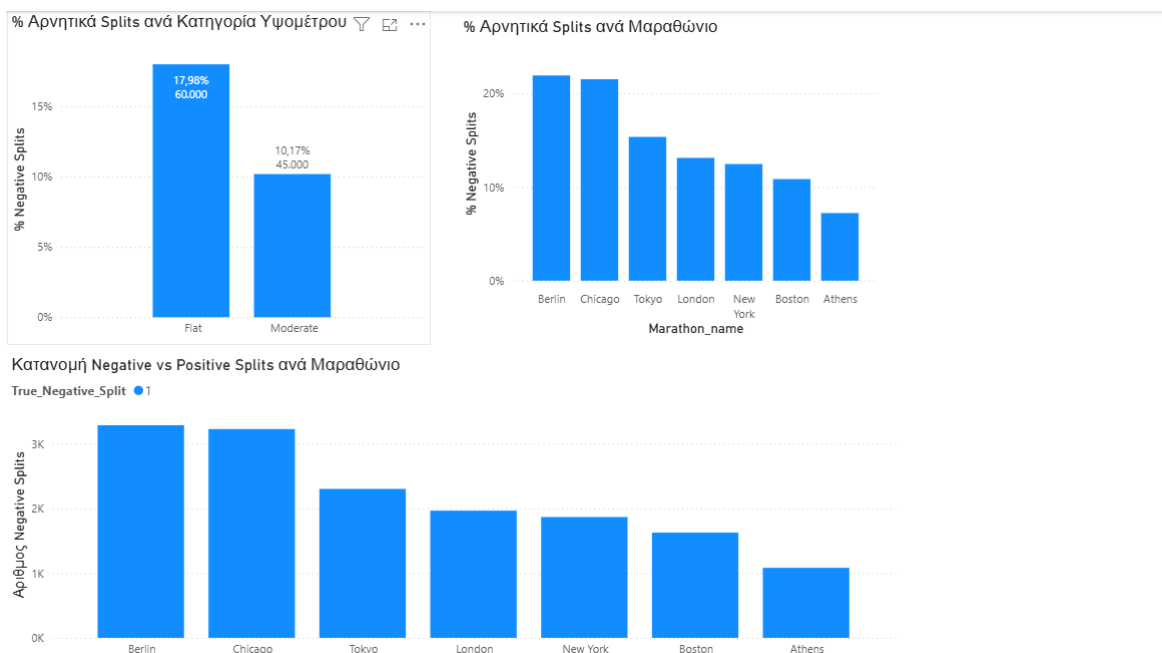
διαδρομών μεσαίας δυσκολίας (10,2%). Το συμπέρασμα επιβεβαιώνει την H3: το γεωγραφικό ανάγλυφο της διαδρομής αποτελεί καθοριστικό παράγοντα για τη δυνατότητα εφαρμογής στρατηγικής αρνητικού split.

Κατηγορία Υψομετρικού	N	Θετικό Split (n)	Αρνητικό Split (n)	% Αρνητικό	χ^2 / p-value
Flat ($\leq 100\text{m}$)	60.000	49.214	10.786	18,0%	
Moderate ($>100\text{m}$)	45.000	40.422	4.578	10,2%	
Chi-Square test	105.000				$\chi^2(1)=1.253,507$, $p<0,001$

Πίνακας 4 Αποτελέσματα Ελέγχου χ^2 — Υψομετρικό Ανάγλυφο και Αρνητικό Split

Αντίθετα, στις διαδρομές με έντονο υψομετρικό προφίλ (Moderate), με κυρίαρχο παράδειγμα τον Αυθεντικό Μαραθώνιο Αθηνών, η συχνότητα εμφάνισης αρνητικού περάσματος κατεγράφη στο 10,2% - σημαντικά χαμηλότερη από το 18,0% των επίπεδων διαδρομών. Η μορφολογία της διαδρομής, η οποία στην περίπτωση της Αθήνας περιλαμβάνει παρατεταμένες ανηφόρες μέχρι το 31ο χιλιόμετρο, επιβάλλει μια αναγκαστική εμβιομηχανική και μεταβολική καταπόνηση που αποτρέπει τη δυνατότητα επιτάχυνσης στο δεύτερο μισό. Το αποτέλεσμα αυτό επιβεβαιώνει πλήρως την υπόθεση H3, αναδεικνύοντας το γεωγραφικό ανάγλυφο ως έναν ανυπέρβλητο ανασταλτικό παράγοντα για την εφαρμογή στρατηγικής negative split.

Στην Εικόνα 8 παρουσιάζεται το ποσοστό των δρομέων που πέτυχαν αρνητικό split (negative split) ανά κατηγορία υψομετρικού αναγλύφου (Flat/Moderate). Το ραβδόγραμμα καταγράφει με σαφήνεια τη διαφορά μεταξύ επίπεδων διαδρομών (18,0%, κυρίως Βερολίνο και Σικάγο) και ανάγλυφων διαδρομών (10,2%, κυρίως Αυθεντικός Μαραθώνιος Αθηνών). Το συμπέρασμα είναι αδιαμφισβήτητο: το υψομετρικό ανάγλυφο λειτουργεί ως αποφασιστικός ανασταλτικός παράγοντας που καθιστά αδύνατη την επιτάχυνση στο δεύτερο μισό, επιβεβαιώνοντας πλήρως την υπόθεση H3 και αναδεικνύοντας το γεωγραφικό ανάγλυφο ως κυρίαρχο καθοριστή της αγωνιστικής στρατηγικής σε διαδρομές μεγάλης υψομετρικής δυσκολίας.



Εικόνα 8 Στιγμιότυπο διαδραστικού πίνακα ελέγχου (Power BI dashboard) – Αρνητικά splits ανά κατηγορία υψομέτρου

8. Συζήτηση

Το παρόν κεφάλαιο ερμηνεύει τα ευρήματα της έρευνας υπό το πρίσμα της διεθνούς βιβλιογραφίας και αξιολογεί τις θεωρητικές και πρακτικές τους προεκτάσεις. Εξετάζεται η φυσιολογική μηχανική που ερμηνεύει κάθε αποτέλεσμα, συγκρίνονται τα ευρήματα με προηγούμενες μελέτες, και αναδεικνύεται ο τρόπος με τον οποίο η ολοκληρωμένη αρχιτεκτονική πληροφοριακού συστήματος μετατρέπει τα ευρήματα σε εφαρμοσμένα εργαλεία αθλητικής αναλυτικής.

8.1 Ερμηνεία των Ευρημάτων και Φυσιολογική Μηχανική

Τα στατιστικά αποτελέσματα που προέκυψαν από την ανάλυση του εκτεταμένου δείγματος των 105.000 δρομέων επιβεβαιώνουν εν μέρει τις ερευνητικές υποθέσεις των πολυπαραγοντικών επιδράσεων που διαμορφώνουν την απόδοση στον μαραθώνιο δρόμο. Η απόρριψη της μηδενικής υπόθεσης στον έλεγχο ANOVA τεκμηριώνει τη φυσιολογική καταπόνηση που προκαλεί το φαινόμενο του θερμικού στρες. Όταν οι δρομείς εκτίθενται στην κατηγορία υψηλών θερμοκρασιών ("Warm"), η αύξηση του Net_time_sec ερμηνεύεται εμβιομηχανικά και αιμοδυναμικά από την ανάγκη του οργανισμού να εκτρέψει μεγάλο όγκο αίματος προς το δέρμα για την αποβολή θερμότητας μέσω εξάτμισης ιδρώτα. Η διαδικασία αυτή, σε συνδυασμό με την προοδευτική αφυδάτωση, μειώνει τον όγκο παλμού της καρδιάς, αναγκάζοντας τον κεντρικό ρυθμιστή (Central Governor) να περιορίσει την επιστράτευση των κινητικών μονάδων των μυών, προκειμένου να αποτραπεί η υπερθερμία.

Αναφορικά με τη διερεύνηση των δημογραφικών παραγόντων, η απουσία στατιστικά σημαντικής διαφοροποίησης στο Pacing_ratio (0,9548 για τους άνδρες έναντι 0,9554 για τις γυναίκες, $p = 0,323$) υποδηλώνει μια συγκλίνουσα φυσιολογική και αγωνιστική συμπεριφορά μεταξύ των δύο φύλων υπό πραγματικές συνθήκες. Από φυσιολογική άποψη, το γεγονός ότι και οι δύο ομάδες παρουσιάζουν μέσο δείκτη ελαφρώς μικρότερο της μονάδας υποδηλώνει μια γενικευμένη τάση για αποτελεσματική διαχείριση δυνάμεων και επίτευξη σταθερού ή ελαφρώς επιταχυνόμενου ρυθμού στο δεύτερο μισό του αγώνα (αρνητικό split). Η ομοιογένεια αυτή αναδεικνύει ότι η προπονητική ωριμότητα, η εμβιομηχανική οικονομία τρεξίματος και η ενεργειακή επάρκεια (μεταβολισμός λιπών και υδατανθράκων) λειτουργούν εξίσου εξισορροπητικά και για τα δύο φύλα στον σύγχρονο

μαζικό αθλητισμό, καταρρίπτοντας τη στερεοτυπική προσέγγιση που ήθελε τους άνδρες δρομείς να υιοθετούν συστηματικά πιο ρισκοκίνδυνες και αστάθετες στρατηγικές ρυθμού (pacing profiles).

8.2 Σύγκριση με τη Διεθνή Βιβλιογραφία

Τα ευρήματα της παρούσας διπλωματικής εργασίας έρχονται σε πλήρη συμφωνία με τις κλασικές μελέτες των Ely et al. (2007, 2008), οι οποίοι είχαν επισημάνει ότι η αύξηση της θερμοκρασίας πάνω από το όριο των 11°C προκαλεί μια προβλέψιμη και βαθμιαία υποβάθμιση των χρόνων τερματισμού. Η στατιστική επιβεβαίωση της ANOVA και των post-hoc ελέγχων Tukey δείχνει ότι η αρνητική επίδραση του κλίματος παραμένει σταθερή και επαληθεύσιμη ακόμα και σε σύγχρονα, διευρυμένα σύνολα δεδομένων που περιλαμβάνουν διαφορετικές γεωγραφικές τοποθεσίες.

Στο πεδίο των στρατηγικών pacing, ο έλεγχος t-test δεν αποκάλυψε στατιστικά σημαντική διαφορά μεταξύ ανδρών και γυναικών [$t(15.370) = -0,988$, $p = 0,323$], γεγονός που δεν επιβεβαιώνει τα ευρήματα των March et al. (2011) και Diaz et al. (2015). Το αποτέλεσμα αυτό υποδηλώνει ότι, στο εκτεταμένο δείγμα των 105.000 δρομέων, δεν υφίσταται στατιστικά τεκμηριωμένη διαφορά στη στρατηγική pacing μεταξύ των δύο φύλων, αμφισβητώντας την άποψη ότι οι άνδρες υιοθετούν συστηματικά πιο ρισκοκίνδυνη αγωνιστική συμπεριφορά.

Τέλος, η ισχυρή συσχέτιση που ανέδειξε ο έλεγχος X^2 μεταξύ της κατηγορίας Elevation_Category και της απουσίας αρνητικών splits συμπίπτει με τις εμπιομηχανικές αναλύσεις των Gottschall και Kram (2005). Η μελέτη μας επιβεβαιώνει ότι σε απαιτητικές διαδρομές με έντονο ανάγλυφο, όπως ο Μαραθώνιος Αθηνών, η επίτευξη αρνητικού split περιορίζεται σε οριακά ποσοστά (Flat: 18,0% έναντι Moderate: 10,2%), καθώς η παρατεταμένη ανηφορική κλίση επιβάλλει μια ανυπέρβλητη ενεργειακή δαπάνη που εξαντλεί πρόωρα τα μεταβολικά αποθέματα.

8.3 Θεωρητικές και Πρακτικές Προεκτάσεις

Η ενσωμάτωση της ανάλυσης δεδομένων σε ένα ολοκληρωμένο πληροφοριακό σύστημα (Python, SQL Server, Power BI) μετατρέπει τα θεωρητικά συμπεράσματα της αθλητικής επιστήμης σε εφαρμοσμένα εργαλεία λήψης αποφάσεων. Οι διαδραστικοί πίνακες ελέγχου που αναπτύχθηκαν στο Power BI επιτρέπουν σε προπονητές και αθλητές να εκτελούν

προσομοιώσεις και να προσαρμόζουν τη στρατηγική τους με βάση τις επικρατούσες συνθήκες.

Για παράδειγμα, εάν η πρόγνωση του καιρού για μια διοργάνωση δείχνει μετάβαση στην κατηγορία "Warm", ο προπονητής μπορεί να χρησιμοποιήσει τα ποσοτικά δεδομένα της έρευνας για να επαναπροσδιορίσει τον στόχο του Pacing_ratio, προτείνοντας έναν πιο συντηρητικό αρχικό ρυθμό (πλησιέστερο στο even pacing) ώστε να μετριαστεί η αναμενόμενη καρδιαγγειακή μετατόπιση.

Επιπλέον, η τεκμηριωμένη επίδραση του υψομετρικού αναγλύφου προσφέρει ένα αντικειμενικό κριτήριο για την περιοδικότητα της προπόνησης. Οι δρομείς που προετοιμάζονται για διαδρομές με προφίλ παρόμοιο με της Αθήνας καθίσταται σαφές ότι πρέπει να δώσουν έμφαση στην ενδυνάμωση για την αντιμετώπιση των έκκεντρων συσπάσεων των κατηφορικών τμημάτων και να αποφύγουν τον σχεδιασμό στρατηγικών negative split, εστιάζοντας αντίθετα στη διαχείριση της προσπάθειας βάσει της κλίσης του εδάφους.

9. Συμπεράσματα

Το παρόν κεφάλαιο συνοψίζει τα κύρια ευρήματα της έρευνας και αξιολογεί τον βαθμό επίτευξης των ερευνητικών στόχων. Παρουσιάζονται οι περιορισμοί που ενδέχεται να επηρεάζουν τη γενίκευση των αποτελεσμάτων, καθώς και συγκεκριμένες προτάσεις για τη μελλοντική επέκταση της έρευνας στο πεδίο των Sports Analytics και της αθλητικής αναλυτικής μεγάλης κλίμακας.

9.1 Σύνοψη Κύριων Ευρημάτων

Η παρούσα διπλωματική εργασία ολοκλήρωσε μια εκτενή ποσοτική και αναδρομική μελέτη με τη χρήση προηγμένων εργαλείων πληροφορικής και αθλητικής αναλυτικής (Python Web Scraping, SQL Server, Power BI), αναδεικνύοντας τη σημασία των Sports Data Analytics στον σύγχρονο αθλητισμό αντοχής. Μέσα από τη στατιστική επεξεργασία του δείγματος των 105.000 δρομέων από τους World Marathon Majors και τον Μαραθώνιο Αθηνών, δόθηκαν σαφείς και επιστημονικά τεκμηριωμένες απαντήσεις στα κεντρικά ερευνητικά ερωτήματα της έρευνας, επιβεβαιώνοντας εν μέρει τις αρχικές ερευνητικές υποθέσεις (H1 και H3 επιβεβαιώθηκαν, H2 απορρίφθηκε).

Αρχικά, αναφορικά με την επίδραση του περιβάλλοντος, η Ανάλυση Διακύμανσης (ANOVA) τεκμηρίωσε ότι η περιβαλλοντική θερμοκρασία ασκεί καθοριστική επίδραση στην απόδοση των αθλητών. Οι υψηλές θερμοκρασίες (κατηγορία Warm) σχετίζονται με στατιστικά σημαντική υποβάθμιση του Performance_Index και αύξηση του συνολικού χρόνου ολοκλήρωσης του αγώνα, αναδεικνύοντας τις επιπτώσεις του θερμικού στρες και της καρδιαγγειακής μετατόπισης.

Δεύτερον, η διερεύνηση των δημογραφικών χαρακτηριστικών μέσω του ελέγχου t-test κατέρριψε την αρχική υπόθεση περί συστηματικής διαφοροποίησης μεταξύ των δύο φύλων ως προς τη στρατηγική κατανομής έργου. Οι άνδρες και οι γυναίκες δρομείς επέδειξαν στατιστικά ισοδύναμη σταθερότητα ρυθμού, με τις μέσες τιμές του δείκτη Pacing_ratio (0,9548 και 0,9554 αντίστοιχα) να μην παρουσιάζουν στατιστικά σημαντική διαφορά ($p = 0,323$). Το εύρημα αυτό υπογραμμίζει ότι η αγωνιστική συμπεριφορά και η ανθεκτικότητα στην κόπωση κατά τα τελευταία στάδια της προσπάθειας διαμορφώνονται από το προπονητικό υπόβαθρο και τη φυσιολογική προετοιμασία του ατόμου, ανεξαρτήτως φύλου.

Τέλος, ο έλεγχος συνάφειας X^2 ανέδειξε το γεωγραφικό ανάγλυφο της διαδρομής ως έναν ανυπέρβλητο παράγοντα για την αγωνιστική στρατηγική, καθώς διαδρομές με έντονες υψομετρικές κλίσεις (κατηγορία Moderate), όπως η Αυθεντική Διαδρομή της Αθήνας, αποτρέπουν συστηματικά την επίτευξη αρνητικού περάσματος (negative split) λόγω της πρόωρης μεταβολικής και εμβιομηχανικής κόπωσης που προκαλεί το παρατεταμένο ανηφορικό τμήμα.

9.2 Περιορισμοί της Έρευνας

Παρά την εγκυρότητα και τη σημασία των συμπερασμάτων, η παρούσα έρευνα υπόκειται σε συγκεκριμένους περιορισμούς που απορρέουν από τη φύση και τη δομή των διαθέσιμων αναδρομικών δεδομένων του αρχείου.

Ο κυριότερος περιορισμός αφορά την αποκλειστική χρήση της θερμοκρασίας ως περιβαλλοντικής μεταβλητής, χωρίς τη δυνατότητα συνυπολογισμού της σχετικής υγρασίας του αέρα, της ηλιακής ακτινοβολίας ή της ταχύτητας του ανέμου, παράμετροι οι οποίοι, σύμφωνα με τη διεθνή βιβλιογραφία, συνδιαμορφώνουν τον δείκτη υγρού θερμομέτρου (WBGT) και επηρεάζουν τη θερμορρύθμιση.

Επιπλέον, η βάση δεδομένων στερούνταν κρίσιμων δημογραφικών και φυσιολογικών στοιχείων των συμμετεχόντων, όπως η ηλικιακή κατηγορία, το προπονητικό υπόβαθρο (εβδομαδιαίος όγκος χιλιομέτρων), το σωματικό βάρος, καθώς και η μέγιστη πρόσληψη οξυγόνου (Vo_{2max}). Η απουσία αυτών των μεταβλητών περιορίζει τη δυνατότητα διεξαγωγής μιας πιο εξειδικευμένης πολυπαραγοντικής ανάλυσης, καθώς η ηλικία και η προπονητική εμπειρία αποτελούν γνωστούς παράγοντες που επηρεάζουν τη σταθερότητα του ρυθμού και την ανθεκτικότητα στην κόπωση.

9.3 Προτάσεις για Μελλοντική Έρευνα

Με βάση τα ευρήματα και τους περιορισμούς της παρούσας διπλωματικής εργασίας, προτείνονται συγκεκριμένες κατευθύνσεις για τη μελλοντική επέκταση της επιστημονικής έρευνας στο πεδίο των Sports Analytics.

Προτείνεται η ενσωμάτωση πιο σύνθετων μετεωρολογικών δεικτών στα μοντέλα ανάλυσης, καθώς και η διεύρυνση του δείγματος με τη συμπερίληψη της ηλικίας των δρομέων,

προκειμένου να μελετηθεί πώς μεταβάλλεται η στρατηγική racing κατά τη διαδικασία της γήρανσης.

Από τεχνολογικής άποψης, η έρευνα μπορεί να επεκταθεί με την εφαρμογή προηγμένων αλγορίθμων Μηχανικής Μάθησης (Machine Learning), όπως τα Τυχαία Δάση (Random Forests) ή τα Νευρωνικά Δίκτυα (Neural Networks). Τα μοντέλα αυτά, τροφοδοτούμενα με real-time δεδομένα από φορετούς αισθητήρες (wearables) και έξυπνα ρολόγια (GPS watches), θα μπορούν να προβλέψουν την πιθανότητα εμφάνισης πρόωρης εξάντλησης ("the wall") και να προτείνουν δυναμικές προσαρμογές του ρυθμού κατά τη διάρκεια του αγώνα.

Τέλος, η διασύνδεση των πληροφοριακών συστημάτων διαχείρισης βάσεων δεδομένων με εφαρμογές κινητών τηλεφώνων μπορεί να προσφέρει στους αθλητές εξατομικευμένα πλάνα αγώνα, λαμβάνοντας υπόψη το τρισδιάστατο υψομετρικό προφίλ της διαδρομής και τις αναμενόμενες κλιματικές συνθήκες, γεφυρώνοντας πλήρως την τεχνολογία των Big Data με την εφαρμοσμένη προπονητική επιστήμη.

Βιβλιογραφία

- Angus, S. D. (2014). An optimal pacing strategy for flat marathon courses under metabolic and thermal constraints. *Medicine & Science in Sports & Exercise*, 46(11), 2130–2139.
- Coyle, E. F., & González-Alonso, J. (2001). Cardiovascular drift during prolonged exercise: New insights into mechanisms. *International Journal of Sports Medicine*, 22(4), 223–232.
- Dehne, F., Kong, Q., Rau-Chaplin, A., Zaboli, H., & Zhou, R. (2015). Scalable real-time OLAP on cloud architectures. *Journal of Parallel and Distributed Computing*, 79-80, 31–41.
- Diaz, J. J., Fernandez-Ozcorta, E. J., & Santos-Concejero, J. (2015). The negative split pacing strategy as a determinant of world-class marathon performance. *International Journal of Sports Physiology and Performance*, 10(6), 712–721.
- Ely, M. R., Cheuvront, S. N., Roberts, W. O., & Montain, S. J. (2007). Impact of ambient temperature on marathon running performance. *Medicine & Science in Sports & Exercise*, 39(3), 526–533.
- Ely, M. R., Martin, D. E., Cheuvront, S. N., & Montain, S. J. (2008). Effect of ambient temperature on elite and recreational marathon performance. *International Journal of Sports Medicine*, 29(11), 915–921.
- Galloway, S. D., & Maughan, R. J. (1997). Effects of ambient temperature on the capacity to perform prolonged cycle exercise in man. *Medicine & Science in Sports & Exercise*, 29(9), 1240–1249.
- Gottschall, J. S., & Kram, R. (2005). Ground reaction forces and muscle activity during uphill and downhill running. *Biomedical Engineering Online*, 4(1), 15–24.
- Helou, P., El Helou, N., Guillaume, M., Nassif, H., Berthelot, G., Bowen, M., Tafflet, M., Hamard, L., Volodina, O., Latiri, I., & Toussaint, J. F. (2012). Impact of environmental parameters on marathon running performance. *PLoS ONE*, 7(5), e37462.
- Joyner, M. J., & Coyle, E. F. (2008). Endurance exercise performance: The physiology of champions. *The Journal of Physiology*, 586(1), 35–44.

- March, D. S., Vanderburgh, P. M., Titlebaum, P. J., & Hoops, M. L. (2011). Age, gender and pacing characteristics of recreational marathon runners. *Journal of Strength and Conditioning Research*, 25(11), 2945–2951.
- Maughan, R. J., Otani, H., & Watson, P. (2007). Influence of climate on pacing in long-distance running. *International Journal of Sports Physiology and Performance*, 2(2), 115–128.
- Maughan, R. J., & Shirreffs, S. M. (2010). Dehydration and rehydration in competitive sport. *Scandinavian Journal of Medicine & Science in Sports*, 20(s3), 40–47.
- Montain, S. J., & Coyle, E. F. (1992). Influence of graded dehydration on hyperthermia and cardiovascular drift during exercise. *Journal of Applied Physiology*, 73(4), 1340–1350.
- Noakes, T. D. (2012). The Central Governor Model of Exercise Regulation: An introduction to the modern concepts. *British Journal of Sports Medicine*, 46(1), 1–3.
- Sproule, J. (1998). Running economy and topographical constraints in elite distance athletes. *Sports Medicine*, 26(5), 311–322.
- Tardío, R., Maté, A., & Trujillo, J. (2020). A new big data benchmark for OLAP cube design using data pre-aggregation techniques. *Applied Sciences*, 10(23), 8674.
- Vaisman, A., & Zimányi, E. (2014). *Data Warehouse Systems: Design and Implementation.*, Springer, 1st Edition.
- Βασιλακόπουλος, Μ. (2018). *Σχεδιασμός και Ανάπτυξη Συστημάτων Διαχείρισης Βάσεων Δεδομένων.*, Έκδοση Ελληνικού Ανοικτού Πανεπιστημίου, 1η Έκδοση.
- Γιαννακόπουλος, Κ., & Μαυρομμάτης, Γ. (2015). *Προχωρημένες Μέθοδοι Στατιστικής Ανάλυσης και Πληροφοριακά Συστήματα Υγείας και Αθλητισμού.*, Έκδοση Ελληνικού Ανοικτού Πανεπιστημίου, 2η Έκδοση.
- Κοτζαμανίδης, Χ. (2019). *Εργοφυσιολογία και Προπονητική Καθοδήγηση σε Αγωνίσματα Αντοχής Μεγάλης Διάρκειας.*, Εκδόσεις Γράφημα, 1η Έκδοση.
- Μαργαρίτης, Κ., & Τζιόλας, Χ. (2017). *Τεχνολογίες Διαχείρισης Μεγάλων Συνόλων Δεδομένων και Επιχειρηματική Ευφυΐα.*, Έκδοση Ελληνικού Ανοικτού Πανεπιστημίου, 1η Έκδοση.

Μαριδάκης, Σ. (2021). *Εφαρμοσμένη Στατιστική στις Αθλητικές Επιστήμες με τη Χρήση Πληροφοριακών Συστημάτων*, Εκδόσεις Πασχαλίδης, 3η Έκδοση.

Παπαδόπουλος, Α. Ν., Τζουραμάνης, Θ., Γούναρης, Α., & Μανωλόπουλος, Ι. (2020). *Συστήματα Βάσεων Δεδομένων: Θεωρία, Πρακτική, Εφαρμογή*, Εκδόσεις Νέων Τεχνολογιών, 2η Έκδοση.

Athens Authentic Marathon - Results. <https://www.athensauthenticmarathon.gr/en/results>.
Ημερομηνία πρόσβασης: 01-04-2026.

Bank of America Chicago Marathon. <https://www.chicagomarathon.com/>. Ημερομηνία πρόσβασης: 02-01-2026.

Boston Marathon Results | Boston Athletic Association. <https://www.baa.org/races/boston-marathon/results/>. Ημερομηνία πρόσβασης: 23-01-2026.

BMW BERLIN-MARATHON - Results. <https://www.bmw-berlin-marathon.com/en/your-race/results>. Ημερομηνία πρόσβασης: 02-04-2026.

TCS New York City Marathon - Race Results. <https://www.nyrr.org/tcsnycmarathon/results/race-results>. Ημερομηνία πρόσβασης: 02-04-2026.

TCS London Marathon 2024 - Leaderboard. <https://results.tcslondonmarathon.com/2024/?pid=leaderboard&pidp=leaderboard>.
Ημερομηνία πρόσβασης: 12-04-2026.

Tokyo Marathon 2024 - Past Events. <https://www.marathon.tokyo/en/about/past/2024/>.
Ημερομηνία πρόσβασης: 30-04-2026.

Abbott World Marathon Majors. <https://www.worldmarathonmajors.com/>. Ημερομηνία πρόσβασης: 12-03-2026.

Παράρτημα Α: Ανάπτυξη κώδικα Web Scraping

Το παράρτημα αυτό περιλαμβάνει τα αρχεία κώδικα που αναπτύχθηκαν για τη διαδικασία εξαγωγής, επεξεργασίας και αποθήκευσης των δεδομένων από τους επτά μααραθωνίους.

Περιέχει στιγμιότυπα από τον κώδικα και σχολιάζεται η λειτουργία κάθε αρχείου.

```
from selenium import webdriver
from selenium.webdriver.common.by import By
from selenium.webdriver.chrome.options import Options
from selenium.webdriver.support.ui import WebDriverWait
from selenium.webdriver.support import expected_conditions as EC
import csv
import time

# Ρυθμίσεις Chrome
options = Options()
options.add_argument("--headless") # Τρέχει χωρίς να ανοίξει
παράθυρο
options.add_argument("--no-sandbox")
options.add_argument("--disable-dev-shm-usage")

# Εκκίνηση browser
driver = webdriver.Chrome(options=options)

url =
"https://www.athensauthenticmarathon.gr/results/home/organisation:Aut
hentic/year:2022/organisation_race:Marathon"
driver.get(url)

# Περιμένει να φορτωθεί ο πίνακας (mexri 15 defterolepta)
WebDriverWait(driver, 15).until(
    EC.presence_of_element_located((By.CSS_SELECTOR, "table"))
)

time.sleep(3) # μικρή καθυστέρηση για ασφάλεια

# Εντοπισμός πίνακα
table = driver.find_element(By.CSS_SELECTOR, "table")

# Λήψη επικεφαλίδων
headers = [th.text.strip() for th in
table.find_elements(By.CSS_SELECTOR, "thead th")]

# Λήψη όλων των γραμμών
rows = []
for tr in table.find_elements(By.CSS_SELECTOR, "tbody tr"):
    cells = [td.text.strip() for td in
tr.find_elements(By.CSS_SELECTOR, "td")]
    if cells:
        rows.append(cells)

# Αποθήκευση σε CSV
filename = "athens_marathon_2022_results.csv"
with open(filename, "w", newline="", encoding="utf-8") as f:
    writer = csv.writer(f)
    writer.writerow(headers)
    writer.writerows(rows)
```

```
print(f" Τα dedomena apothikeftikan sto '{filename}' ({len(rows)}  
grammes).")
```

```
driver.quit()
```

Αντίστοιχα έχουν γίνει και οι υπόλοιπες κλήσεις για τους μαραθωνίους του Λονδίνου, Βοστώνης, Τόκιο, Βερολίνου, Σικάγο και Νέα Υόρκη για τα έτη 2022, 2023 και 2024.

Παράρτημα Β: Κώδικας Στατιστικής Ανάλυσης

Το παράρτημα αυτό περιλαμβάνει στιγμιότυπα του κώδικα στατιστικής ανάλυσης που αναπτύχθηκε για την επαλήθευση των τριών ερευνητικών υποθέσεων. Οι στατιστικοί έλεγχοι (ANOVA, t-test, Chi-Square) εφαρμόστηκαν στο σύνολο των 105.000 εγγραφών του All_Marathons.csv με χρήση της γλώσσας Python και των βιβλιοθηκών pandas, scipy και pingouin.

B.1 statistical_analysis.py - Φόρτωση δεδομένων και περιγραφική στατιστική

Αρχικά φορτώνεται το ενοποιημένο αρχείο All_Marathons.csv και υπολογίζεται η περιγραφική στατιστική για τις κύριες μεταβλητές (καθαρός χρόνος, Pacing_ratio, θερμοκρασία, υψομετρική άνοδος). Παρακάτω φαίνεται στιγμιότυπο της φόρτωσης και εκτύπωσης του Πίνακα 7.1.

```
df = pd.read_csv('All_Marathons.csv')
df['True_Negative_Split'] = (df['Pacing_ratio'] <= 1.0).astype(int)
df['Net_time_min'] = df['Net_time_sec'] / 60

vars_q = ['Net_time_sec', 'Pacing_ratio',
          'Performance_Index', 'Temp', 'total_elevation_gain']
for v in vars_q:
    s = df[v]
    print(f'{v}: N={s.count()}, M={s.mean():.3f},
          SD={s.std():.3f}, Med={s.median():.2f}')
```

B.2 statistical_analysis.py - ANOVA: Επίδραση Θερμοκρασίας (H1)

Εφαρμογή one-way ANOVA για τον έλεγχο της υπόθεσης H1 (η θερμοκρασία επηρεάζει στατιστικά σημαντικά τον χρόνο τερματισμού). Οι ομάδες σύγκρισης ορίστηκαν ως Cold ($\leq 10^{\circ}\text{C}$), Mild ($11\text{--}18^{\circ}\text{C}$) και Warm ($>18^{\circ}\text{C}$). Παρακάτω φαίνεται στιγμιότυπο του ελέγχου ANOVA με post-hoc Tukey.

```
from scipy.stats import f_oneway
from statsmodels.stats.multicomp import pairwise_tukeyhsd

groups = [df[df['Temp_Category']==cat]['Net_time_sec']
          for cat in ['Cold', 'Mild', 'Warm']]
F, p = f_oneway(*groups)
print(f'ANOVA: F={F:.3f}, p={p:.4f}')

# Post-hoc Tukey
tukey = pairwise_tukeyhsd(
    df['Net_time_sec'], df['Temp_Category'], alpha=0.05)
print(tukey.summary())
```

B.3 statistical_analysis.py — t-test: Φύλο και Pacing_ratio (H2)

Εφαρμογή ανεξάρτητου t-test για τον έλεγχο της υπόθεσης H2 (οι γυναίκες διατηρούν σταθερότερο ρυθμό από τους άνδρες). Η ανάλυση πραγματοποιήθηκε στις 105.000 εγγραφές. Παρακάτω φαίνεται στιγμιότυπο του ελέγχου.

```
from scipy.stats import ttest_ind

df_pacing = df[df['Pacing_ratio'].notna()].copy()
males     = df_pacing[df_pacing['Gender']=='Male']['Pacing_ratio']
females   = df_pacing[df_pacing['Gender']=='Female']['Pacing_ratio']

t, p = ttest_ind(males, females, equal_var=True) # Standard t-test
print(f't({len(df_pacing)-2}) = {t:.3f}, p = {p:.3f}')
print(f'Males M={males.mean():.4f}, Females M={females.mean():.4f}')
```

B.4 statistical_analysis.py — Chi-Square: Ανάγλυφο και Αρνητικό Split (H3)

Εφαρμογή ελέγχου Chi-Square (χ^2) για τον έλεγχο της υπόθεσης H3 (η υψομετρική κατηγορία επηρεάζει σημαντικά την πιθανότητα επίτευξης αρνητικού split). Δημιουργήθηκε πίνακας συνάφειας 2×2 (Elevation_Category × True_Negative_Split). Παρακάτω φαίνεται στιγμιότυπο του ελέγχου.

```
from scipy.stats import chi2_contingency

ct = pd.crosstab(
    df['Elevation_Category'],
    df['True_Negative_Split']
)
chi2, p, dof, expected = chi2_contingency(ct)
n = ct.values.sum()
cramers_v = (chi2 / (n * (min(ct.shape) - 1))) ** 0.5

print(f'chi2({dof}) = {chi2:.3f}, p = {p:.4f}')
print(f'Cramers V = {cramers_v:.4f}')
print(ct)
```

Υπεύθυνη Δήλωση Συγγραφέα:

Δηλώνω ρητά ότι, σύμφωνα με το άρθρο 8 του Ν.1599/1986, η παρούσα εργασία αποτελεί αποκλειστικά προϊόν προσωπικής μου εργασίας, δεν προσβάλλει κάθε μορφής δικαιώματα διανοητικής ιδιοκτησίας, προσωπικότητας και προσωπικών δεδομένων τρίτων, δεν περιέχει έργα/εισφορές τρίτων για τα οποία απαιτείται άδεια των δημιουργών/δικαιούχων και δεν είναι προϊόν μερικής ή ολικής αντιγραφής, οι πηγές δε που χρησιμοποιήθηκαν περιορίζονται στις βιβλιογραφικές αναφορές και μόνον και πληρούν τους κανόνες της επιστημονικής παράθεσης.