Εθνικο Μετσοβιο Πολυτεχνειο

Σχολη Ηλεκτρολογων Μηχανικων & Μηχανικων Υπολογιστων

Εξόρυξη Γνώσης από Δεδομένα, Χειμερινό Εξάμηνο 2019-2020

Εξαμηνιαία Εργασία: ***London Bicycle Hires 2016***

Καράτσαλος Χρήστος, ΔΠΜΣ $\varepsilon.\delta\varepsilon.\mu^2$

Ράπτης Παναγιώτης, ΔΠΜΣ $\varepsilon.\delta\varepsilon.\mu^2$

Αθήνα, 2020

Περιεχόμενα

1 Εισαγωγή **2**
1.1 Περιγραφή Προβλήματος . . . 2
1.2 Συλλογή Δεδομένων - Datasets . . . 3
1.3 Σχετική Δουλειά - Προγενέστερες Προσεγγίσεις . . . 5

2 Εξερεύνηση Δεδομένων **6**
2.1 Προεπεξεργασία Δεδομένων . . . 7
2.2 Δημιουργία Βοηθητικών Χαρακτηριστικών (Feature Generation) . . . 8
2.3 Εξερεύνηση Δεδομένων (Data Exploration) . . . 10

3 Τεχνικές Μηχανικής Μάθησης **19**
3.1 Πρόβλεψη (Prediction) . . . 19
3.1.1 Random Forest (RF) Regressor . . . 20
3.1.2 Τεχνικές Βαθιάς Μάθησης (GRU, LSTM, BiLSTM) . . . 21
3.2 Συσταδοποίηση (Clustering) . . . 23
3.2.1 K-Means Μέθοδος . . . 23
3.2.2 DBSCAN Μέθοδος . . . 25

4 Σύνοψη Αποτελεσμάτων & Μελλοντικές Κατευθύνσεις **26**
4.1 Σύνοψη Αποτελεσμάτων & Σχολιασμός . . . 26
4.2 Μελλοντικές Κατευθύνσεις . . . 28

1 Εισαγωγή

Στο πλαίσιο της συγκεκριμένης εργασίας, καλούμαστε να εξορύξουμε όσο το δυνατόν περισσότερη γνώση από πραγματικά δεδομένα, εστιάζοντας σε ένα συγκεκριμένο πρόβλημα. Έχοντας ως γνώμονα το έντονο ενδιαφέρον μας για την προώθηση εναλλακτικών τρόπων μετακίνησης, εντός πυκνοκατοικημένων μεγαλουπόλεων, αποφασίσαμε να ασχοληθούμε με το dataset *London Bicycle Hires* [1], το οποίο προσφέρεται από το *Google Cloud Platform* και αφορά όλες τις ενοικιάσεις ποδηλάτων που πραγματοποιήθηκαν στο Λονδίνο για τα έτη 2015, 2016 και τους πρώτους επτά μήνες του 2017. Φυσικά, λόγω προβλημάτων με την διαθέσιμη μνήμη RAM, καθώς και τον χρόνο εκτέλεσης ορισμένων λειτουργιών, αποφασίσαμε να ασχοληθούμε μόνο με το έτος 2016, λαμβάνοντας υπόψιν μας τα παρόμοια patterns που παρουσιάζονται, στο σύνολο τους, μεταξύ των τριών αυτών χρόνων.

Έπειτα από *κατάλληλη προεπεξεργασία* του κύριου dataset, λαμβάνοντας υπόψη το μικρό πλήθος αξιόπιστων χαρακτηριστικών προς εκμετάλλευση που μας παρέχει, κρίναμε αναγκαία την *συνένωση* με δύο επιπλέον datasets, σχετικά με τις καιρικές συνθήκες που επικρατούσαν στο Λονδίνο, καθώς και κατάταξη του εκάστοτε σταθμού ενοικίασης ποδηλάτων στην κατάλληλη ζώνη, υιοθετώντας την ζώνη στην οποία ανήκει ο πλησιέστερος σταθμός μετρό. Εν συνεχεία, πραγματοποιείται η *εξερεύνηση* του συνδυασμένου συνόλου δεδομένων, παρουσιάζοντας ορισμένα χρήσιμα στατιστικά, οπτικοποιώντας καταλλήλως την υφιστάμενη πληροφορία και παραθέτοντας τον απαιτούμενο σχολιασμό. Έπειτα, όσον αφορά το κομμάτι των *τεχνικών μηχανικής μάθησης*, επικεντρωνόμαστε στην *πρόβλεψη* (*prediction*) της ωριαίας ζήτησης, καθώς και της έγκυρης *συσταδοποίηση* (*clustering*) των σταθμών ενοικίασης, με βάση την τοποθεσία τους. Τέλος ακολουθεί η παράθεση όλων των βασικών συμπερασμάτων που καταλήξαμε, στο πλαίσιο της συγκεκριμένης εργασίας, καθώς επίσης προτείνονται και ορισμένες μελλοντικές κατευθύνσεις για περαιτέρω και πιο βαθιά επισκόπηση του συγκεκριμένου πεδίου.

1.1 Περιγραφή Προβλήματος

Τα τελευταία χρόνια, τα συστήματα *ενοικίασης ποδηλάτων* (*Bicycle Sharing Systems - BSS*) έχουν γίνει ιδιαίτερα δημοφιλή για την μετακίνηση εντός μεγάλων πόλεων, καθώς καλύπτουν τις αδυναμίες των υφιστάμενων δικτύων δημόσιας συγκοινωνίας. Ταυτόχρονα αποτελούν ένα οικονομικό, οικολογικό, αποδοτικό αλλά και υγιεινό εναλλακτικό μέσο μεταφοράς, που πλέον εντοπίζεται σε πάρα πολλές πόλεις στην Ευρώπη, την Ασία, την Αυστραλία και την Αμερική. Μεταξύ των κορυφαίων *BSS* συγκαταλέγονται εκείνα του Παρισιού, της Νέας Υόρκης, της Σαγκάη, του Λονδίνου, της Βαρκελώνης, του Μόντρεαλ και πολλά ακόμα.

Εστιάζοντας την προσοχή μας σε *συστήματα ενοικίασης ποδηλάτων* μεγάλου μεγέθους, τα οποία απαρτίζονται από πολυάριθμους σταθμούς και ποδήλατα, σε συνδυασμό με την δεδομένη εμπλοκή αρκετών εταιριών, γίνεται εμφανές ότι προκειμένου οι πελάτες να είναι ικανοποιημένοι, επιβάλλεται να εξασφαλιστεί, κατ' ελάχιστον, η υψηλή διαθεσιμότητα ποδηλάτων σε κεντρικούς σταθμούς. Στην πραγματικότητα, όμως, η επίτευξη ενός τέτοιου στόχου κάθε άλλο παρά τετριμμένη είναι, δεδομένου ότι οι διαδρομές των πελατών είναι χρονικά μεταβαλλόμενες και δυναμικές, ενώ επιπρόσθετα η *αναδιανομή* (*redistributing*) των ποδηλάτων είναι μία σχετικά ακριβή διαδικασία για τις εταιρίες ενοικίασης (*bicycle vendors*). Προβάλλει, επομένως, επιτακτική η ανάγκη εξαντλητικής μελέτης και προσεκτικής επιλογής του τρόπου με βάση τον οποίο θα πραγματοποιηθεί το απαιτούμενο *rebalancing*.

Figure 1: Σταθμός ενοικίασης ποδηλάτων της γνωστής εταιρείας CitiBike στην Νέα Υόρκη

Απαραίτητη προϋπόθεση, φυσικά, είναι η πλήρης κατανόηση των παραγόντων που επηρεάζουν την ζήτηση ποδηλάτων από την πλευρά των πελατών, όπως υφιστάμενα καθημερινά patterns μετακίνησης, σε συγκεκριμένα διαστήματα εντός της ημέρας, συναρτήσει των καιρικών συνθηκών που επικρατούν, της κίνησης στους δρόμους ή στα μέσα μαζικής μεταφοράς, για παράδειγμα, ενώ επίσης ιδιαίτερα χρήσιμη μπορεί να αποδειχθεί και η τοποθεσία του εκάστοτε σταθμού.

Ενστερνιζόμενοι το γενικότερο πνεύμα του συγκεκριμένου μαθήματος, η εργασία έχει δομηθεί με τρόπο ώστε να εστιάζει αποκλειστικά στην εξερεύνηση, τον εντοπισμό και εν τέλει τον χαρακτηρισμό εκείνων των παραγόντων, οι οποίοι συμβάλουν στην διαμόρφωση της ζήτησης από την πλευρά των πελατών. Βασιζόμενοι, λοιπόν, στα πραγματικά δεδομένα που έχουμε στην διάθεση μας, η συγκεκριμένη εργασία, θα μπορούσε κανείς να πει, ότι λειτουργεί ως προάγγελος για την μελέτη του προβλήματος *rebalancing* μελλοντικά, χωρίς όμως να μας απασχολεί καθόλου στο σύνολο αυτής.

Εν συνεχεία παρουσιάζονται συνοπτικά τα σύνολα δεδομένων που χρησιμοποιήσαμε, με την σύμπτυξη των οποίων προέκυψε το τελικό στοχευμένο dataset, στο οποίο πραγματοποιείται η διαδικασία *εξερεύνησης των δεδομένων* (*data exploration*) - βλ. *Κεφάλαιο 2* - ενώ εφαρμόζονται εν συνεχεία ορισμένες *τεχνικές μηχανικής μάθησης* - βλ. *Κεφάλαιο 3*. Όσον αφορά το συγκεκριμένο εισαγωγικό κεφάλαιο, επίσης, παρουσιάζονται εν τέλει συνοπτικά οι κύριες προσεγγίσεις που εντοπίστηκαν στην βιβλιογραφία, σχετικά με *BSS*, στοιχεία από ορισμένες μάλιστα, αποτέλεσαν και πηγή έμπνευσής μας.

1.2 Συλλογή Δεδομένων - Datasets

Όπως προαναφέρθηκε, το κύριο σύνολο δεδομένων που χρησιμοποιούμε είναι το *London Bicycle Hires* [1], το οποίο εμπεριέχει όλες τις ενοικιάσεις ποδηλάτων που πραγματοποιήθηκαν στο Λονδίνο από το 2015 έως και τα μέσα του 2017. Ειδικότερα, το συγκεκριμένο dataset αποτελείται από δύο επιμέρους αρχεία, τα οποία αντιστοιχούν στις ενοικιάσεις (*bike_hires.csv*) καθώς και στους υφιστάμενους σταθμούς ενοικίασης (*bike_stations.csv*), αντίστοιχα. Δυστυχώς, στο σύνολο τους, τα δύο προαναφερθέντα αρχεία δεν συνοδεύονται από κάποιο documentation των χαρακτηριστικών που διαθέτουν, καθιστώντας ορισμένα από τα πεδία τους μη κατανοητά ως προς το τι πρεσβεύουν. Ταυτόχρονα, ορισμένα πεδία αποτελούνται,

στην συντριπτική τους πλειοψηφία, από *missing values*. Ως αποτέλεσμα, αναγκαζόμαστε να απομακρύνουμε τα "παθολογικά" αυτά χαρακτηριστικά των πρωταρχικών συνόλων δεδομένων[1], για το υπόλοιπο της εργασίας, επιτυγχάνοντας ταυτόχρονα ακουσίως, σημαντική μείωση της διάστασης. Φυσικά, τέτοιου είδους δυσκολίες είναι αναμενόμενες, αφού ασχολούμαστε με πραγματικά δεδομένα, τα οποία καλούμαστε να παραλάβουμε και εν συνεχεία να επεξεργαστούμε, με τον καταλληλότερο για εμάς τρόπο, έχοντας ως στόχο την εξαγωγή όσο δυνατόν περισσότερης γνώσης.

```
Mounted at /content/drive
```

	rental_id	bike_id	duration	start_date	start_station_name	start_station_id	end_date	end_station_name	end_station_id
0	40346508	12019	360	2015-01-04 00:00:00 UTC	Harriet Street, Knightsbridge	368	2015-01-04 00:06:00 UTC	Ebury Bridge, Pimlico	424
1	40346509	13032	660	2015-01-04 00:00:00 UTC	Brushfield Street, Liverpool Street	251	2015-01-04 00:11:00 UTC	Regent's Row , Haggerston	553

Figure 2: Δύο πρώτες εγγραφές του bike_hires.csv αρχείου

Πιο αναλυτικά, το αρχείο με τις ενοικιάσεις ποδηλάτων (*bike_hires.csv*) αποτελείται από 24.369.201 διαδρομές συνολικά. Κάθε εγγραφή χαρακτηρίζεται από τον μοναδικό id διαδρομής[2], την διάρκεια διαδρομής[3], το μοναδικό id ποδηλάτου, την ημερομηνία έναρξης και τερματισμού της διαδρομής, καθώς και το όνομα και το id των σταθμών αφετηρίας και προορισμού. Παραπάνω, παρουσιάζονται ενδεικτικά οι δύο πρώτες πλειάδες του συγκεκριμένου αρχείου (βλ. *Figure 2*).

Όσον αφορά το αρχείο με τους σταθμούς ενοικίασης (*bike_stations.csv*), εμπεριέχει 785 διαφορετικούς σταθμούς συνολικά. Σε αυτό το σημείο, ως μία σύντομη παρένθεση, αξίζει να επισημανθεί ότι, δυστυχώς, μεταξύ των διαθέσιμων διαδρομών που εκτελέστηκαν και των σταθμών δεν υπάρχει 1-1 αντιστοίχηση, μιας και διαπιστώσαμε ορισμένες "παθολογικές" περιπτώσεις, όπως για παράδειγμα σταθμοί να έχουν καταργηθεί εντός της τριετίας 2015 με 2017, με το χαρακτηριστικό id τους να ανατίθεται, σε μεταγενέστερο χρόνο, σε κάποιον νεότερο σταθμό. Το ζήτημα αυτό, καθώς και η αντιμετώπιση του συζητούνται αναλυτικά στο επόμενο κεφάλαιο, στο πλαίσιο της φάσης *προεπεξεργασίας* των δεδομένων. Εστιάζοντας και πάλι την προσοχή μας στο αρχείο που περιλαμβάνει τους σταθμούς ενοικίασης, κάθε εγγραφή αποτελείται από το id σταθμού, το γεωγραφικό μήκος και πλάτος του σταθμού, το όνομα του σταθμού, την ημερομηνία εγκατάστασης και απομάκρυνσης του σταθμού, καθώς επίσης και boolean μεταβλητές σχετικά με το εάν ο εκάστοτε σταθμός είναι locked ή/και προσωρινός. Εν συνεχεία παρουσιάζονται ενδεικτικά οι δύο πρώτες πλειάδες αυτού του αρχείου (βλ. *Figure 3*). Όπως επισημάνθηκε και παραπάνω, και στα δύο αρχεία, κρατώντας μόνο εκείνα που κρίναμε ότι μας παρείχαν αξιόπιστη, ως ένα βαθμό, κατανοητή και απαραίτητη πληροφορία για την περαιτέρω ανάλυση που ακολουθεί.

	id	installed	latitude	locked	longitude	name	bikes_count	docks_count	nbEmptyDocks	temporary	terminal_name	install_date	removal_date
0	211	True	51.494645	false	-0.158106	Cadogan Place, Knightsbridge	0	0	0	False	3469	2010-07-19	None
1	250	True	51.489932	false	-0.162727	Royal Avenue 1, Chelsea	0	10	9	False	3440	2010-07-20	None

Figure 3: Δύο πρώτες εγγραφές του bike_stations.csv αρχείου

[1]Για περισσότερες πληροφορίες σχετικά με βασικά στατιστικά στοιχεία των χαρακτηριστικών των αρχικών συνόλων δεδομένων, τα οποία έπαιξαν πραγματικά καθοριστικό ρόλο για την κατανόηση του περιεχομένου των datasets, σε πρώτη φάση τουλάχιστον, σας παραπέμπουμε στο επισυναπτόμενο Python Notebook, το οποίο περιλαμβάνει όλο τον κώδικα Python για το συγκεκριμένο project.

[2]Χρησιμοποιώντας ορολογία βάσεων δεδομένων, το *rental_id* αποτελεί πρωτεύον κλειδί για τις εγγραφές του συγκεκριμένου αρχείου.

[3]Η διάρκεια είναι μετρούμενη σε secs.

Επιπρόσθετα, έχοντας ως στόχο την πραγματοποίηση μίας πιο σφαιρικής μελέτης, σχετικά με το πρόβλημα ενοικίασης ποδηλάτων στο Λονδίνο, προέβαλε επιτακτική η ενσωμάτωση πρόσθετης χρήσιμης πληροφορίας στο αρχικό σύνολο δεδομένων, εξαιτίας των λιγοστών διαθέσιμων χαρακτηριστικών. Ως αποτέλεσμα, σε πρώτη φάση, αποφασίσαμε να εισάγουμε δεδομένα σχετικά με τις καιρικές συνθήκες της εκάστοτε διαδρομής. Πιο αναλυτικά, με την χρήση του *Weather API* που προσφέρει το [2], καταφέραμε να λάβουμε δεδομένα καιρού, ανά ώρα. Ειδικότερα, σε κάθε διαδρομή ενσωματώθηκε πληροφορία σχετικά με την θερμοκρασία εδάφους, την ατμοσφαιρική πίεση, τα επίπεδα υγρασίας, η ένταση του ανέμου που επικρατούσε, καθώς επίσης και κάποια abstract περιγραφή καιρού[4] για την συγκεκριμένη ώρα. Παρομοίως με τα προηγούμενα αρχεία, έπονται οι δύο πρώτες πλειάδες του συγκεκριμένου αρχείου (βλ. *Figure 4*).

	dt_iso	temp	pressure	humidity	wind_speed	weather_main
0	2015-01-01 00:00:00 +0000 UTC	4.78	1033.2	84	4.1	Clouds
1	2015-01-01 01:00:00 +0000 UTC	4.98	1032.9	85	4.1	Clouds

Figure 4: Δύο πρώτες εγγραφές του london_weather.csv αρχείου

Τέλος, κρίθηκε αναγκαία και η συμπερίληψη πληροφορίας σχετικά με τον βαθμό όπου κάποιος σταθμός ενοικίασης ποδηλάτων πλαισιώνεται από το δίκτυο των μέσων μαζικής μεταφοράς του Λονδίνου. Προς αυτή την κατεύθυνση, λαμβάνοντας υπόψη το διαχρονικά εξαιρετικό δίκτυο υπόγειου σιδηρόδρομου που διαθέτει το Λονδίνο, επιλέξαμε κάθε σταθμός ενοικίασης να χαρακτηρίζεται ποιοτικά από το κατά πόσο απέχει από τον πλησιέστερο σταθμό μετρό. Επιπλέον, κάθε σταθμός "βαπτίζεται" με την ζώνη αυτού, επιτυγχάνοντας με αυτόν τον τρόπο τον διαχωρισμό τους σε αντίστοιχες ζώνες με εκείνες των σταθμών του μετρό. Φαινομενικά μία τέτοια προσέγγιση φαντάζει ως αρκετά απλή, παρόλα αυτά στην πράξη αποδεικνύεται ιδιαίτερα αξιόπιστη, προσθέτοντας σημαντική πληροφορία κατά την φάση της *εξερεύνησης των δεδομένων* (βλ. *Κεφάλαιο 2*). Ενδεικτικά, παραθέτονται ακολούθως οι δύο πρώτες πλειάδες του αρχείου που περιλαμβάνει τους σταθμούς του μετρό του Λονδίνου (βλ. *Figure 5*).

	latitude	longitude	zone
0	51.531952	0.003723	3
1	51.490784	0.120272	4

Figure 5: Δύο πρώτες εγγραφές του subway_stations.csv αρχείου

1.3 Σχετική Δουλειά - Προγενέστερες Προσεγγίσεις

Η εκτενής επισκόπηση της υπάρχουσας βιβλιογραφίας, σχετικά με *BSS*, αποτέλεσε την βάση κατά την δόμηση και τον σχεδιασμό της παρούσας εργασίας, αντλώντας αρκετές ενδιαφέρουσες ιδέες, αλλά και κατανοώντας καλύτερα την φύση των σχετικών προβλημάτων που αναδύονται. Στην συγκεκριμένη υποενότητα, πραγματοποιείται μία σύντομη επισκόπηση της υπάρχουσας βιβλιογραφίας για *συστήματα ενοικίασης ποδηλάτων*.

[4] Το πεδίο της περιγραφής του καιρού πρόκειται για μία κατηγορική μεταβλητή, η οποία λαμβάνει τις εξής τιμές: Clear, Clouds, Fog, Mist, Rain, Drizzle, Haze, Snow, Thunderstorm.

Ίσως από τις πρώτες εργασίες προς την κατεύθυνση της αξιοποίησης του *Data Mining* για προβλήματα σχετικά με *BSS*, αποτελεί το [12]. Ειδικότερα, το συγκεκριμένο άρθρο προσφέρει ένα πολύ καθορισμένο και ξεκάθαρο μονοπάτι για εξόρυξη γνώσης, έχοντας ως στόχο την κατανόηση των *activity patterns*, προκειμένου να εντοπιστούν πιθανά *imbalancies* στην κατανομή των διαθέσιμων ποδηλάτων στους σταθμούς. Οφείλουμε να παραδεχτούμε, εάν και πρόκειται για πρωτόλειο, σημαντικό μέρος της ανάλυσης που ακολουθήσαμε, η οποία παρουσιάζεται αναλυτικά στα κεφάλαια που ακολουθούν, βασίστηκε στην συγκεκριμένη εργασία. Προς έκπληξή μας, μελετώντας μεταγενέστερες εργασίες επί του θέματος, εντοπίσαμε το συγκεκριμένο άρθρο αρκετά συχνά στις παραπομπές, αποτελώντας κατά κάποιο τρόπο ένα από τα βασικά θεμέλια της συγκεκριμένης ερευνητικής περιοχής.

Εν συνεχεία παρουσιάζονται ορισμένα άρθρα που υπέπεσαν της προσοχής μας και θεωρούμε άξια σύντομης αναφοράς. Ξεκινώντας, αναφέρουμε το [8] στο οποίο γίνεται μία προσπάθεια διατήρησης του δικτύου όσο το δυνατό πιο *balanced* κατά τις ώρες αιχμής, πραγματοποιώντας το *rebalancing* τις βραδυνές ώρες. Για τον σκοπό αυτό, το συγκεκριμένο πρόβλημα μοντελοποιείται ως ένα πρόβλημα βελτιστοποίησης, για το οποίο γίνεται εντατική προσπάθεια επίλυσης, συγκρίνοντας την απόδοση δύο διαφορετικών μεθόδων. Η ίδια ερευνητική ομάδα, έπειτα από λίγο καιρό, δημοσιεύει το [10] επικεντρωνόμενη στην πρόβλεψη του σταθμού αφετηρίας και προορισμού κατά τις πρωινές ώρες αιχμής τις καθημερινές, χρησιμοποιώντας μια πληθώρα από features. Επίσης, το [13] πραγματεύεται αναλυτικά την *εξόρυξη γνώσης* δεδομένων σχετικών με την ενοικίαση ποδηλάτων, καταλήγοντας στην συνέχεια στην ανάπτυξη ενός μοντέλου για πρόβλεψη διαδρομής. Ας σημειωθεί σε αυτό το σημείο, ότι το πρώτο κομμάτι του συγκεκριμένου paper μας παρείχε αρκετά και ιδιαίτερα χρήσιμα ερεθίσματα, επιρεάζοντας σε μεγάλο βαθμό ορισμένα σημεία της περαιτέρω ανάλυσης που αναπτύσσουμε. Προς έκπληξή μας, αναζητώντας την σχετική βιβλιογραφία, διαπιστώσαμε πληθώρα πρόσθετων διαφορετικών κατευθύνσεων για προβλήματα *BSS*, όπως για παράδειγμα η δυναμική συσταδοποίηση σταθμών ([4]), traffic prediction ([5]), ενώ σημαντική προσπάθεια επιτελείται, λαμβάνοντας πρωτότυπες προσεγγίσεις, για το πρόβλημα του *rebalancing* ([11], [7]).

Καθίσταται επομένως φανερό, ότι προβλήματα που σχετίζονται με *BSS* έχουν απασχολήσει και συνεχίζουν να απασχολούν έντονα την επιστημονική κοινότητα, σε μια προσπάθεια για επίλυση των υπαρκτών δυσκολιών, εξαιτίας της υψηλής πολυπλοκότητας που χαρακτηρίζει τέτοιου είδους προβλήματα.

2 Εξερεύνηση Δεδομένων

Ως γνωστόν, η *εξερεύνηση δεδομένων* αποτελεί το κομμάτι κατά το οποίο απαιτείται η εις βάθος κατανόηση του περιεχομένου των διαθέσιμων δεδομένων, με βάση χρήσιμα στατιστικά στοιχεία, η μελέτη της δομής και των υφιστάμενων συσχετίσεων που χαρακτηρίζουν τα datasets, καθώς και διαφόρων ειδών οπτικοποιήσεις. Προκειμένου να είμαστε σε θέση να εξάγουμε όσο το δυνατό περισσότερη γνώση, είναι αναγκαίο τις περισσότερες φορές να δημιουργηθούν ορισμένα πρόσθετα χαρακτηριστικά (*feature engineering - generation*), με βάση εκείνα που έχουμε στην διάθεση μας. Φυσικά, πρίν από όλη αυτή την διαδικασία προβάλει επιτακτική η ανάγκη για κάποιου είδους "φιλτράρισμα" των δεδομένων, γνωστή και ως *φάση προεπεξεργασίας*, σύμφωνα με την οποία πραγματοποιείται διαχείριση των ελλιπών δεδομένων, διορθώνονται τυχόν ασυνέπειες και μειώνεται η διάσταση των δεδομένων, προετοιμάζοντας το έδαφος για την φάση της *εξερεύνησης των δεδομένων* που ακολουθεί.

Στόχος, επομένως, του παρόντος κεφαλαίου αποτελεί η διερευνητική ανάλυση του συνόλου δεδομένων, με στόχο τον εντοπισμό και κατανόηση όλης της δυνατής γνώσης που μπορεί να εξαχθεί. Έπειτα από την ολοκλήρωση της φάσης *εξερεύνησης των δεδομένων*, ακολουθεί η εφαρμογή διαφόρων *τεχνικών μηχανικής μάθησης* (βλ. *Κεφάλαιο 3*), με κύριο σκοπό την ανακάλυψη μη φανερών patterns στα δεδομένα, τα οποία θα αξιολογηθούν διαισθητικά, με βάση την γνώση και την γενική διαίσθηση που καλλιεργείται στο παρόν κεφάλαιο, μιας και ιδανικά στόχος μας είναι η ανακάλυψη μη προφανών, κατανοητών και χρήσιμων συμπερασμάτων.

2.1 Προεπεξεργασία Δεδομένων

Όπως έχει ήδη υπονοηθεί προηγουμένως, με βάση τα *Figures 2 - 3*, από τα αρχικά datasets έχουν διατηρηθεί μόνο εκείνα τα columns που δεν διέθεταν *missing values*, καθώς επίσης και εκείνα που ήταν κατανοητό τι αντιπροσωπεύουν. Ακολούθως παρουσιάζονται τα *missing values* για τα δύο αρχεία του πρωταρχικού συνόλου δεδομένων, *bike_hires.csv* και *bike_stations.csv*, αντίστοιχα. Στο πλαίσιο της συγκεκριμένης εργασίας, αποφασίστηκε να απομακρυνθούν όλες οι εγγραφές που εμπεριέχουν *missing values*, δεδομένου ότι εμπλέκονται κατηγορικά και όχι αριθμητικά δεδομένα. Ως αποτέλεσμα, δεν καθίσταται δυνατή η εφαρμογή γνωστών μεθόδων από την βιβλιογραφία.[5]. Επιπλέον, απορρίπτονται διαδρομές με σταθμούς αφετηρίας ή προορισμού που δεν υπάρχουν στο αρχείο με τους σταθμούς ενοικιάσεις ποδηλάτων (*bike_stations.csv*), αφού θεωρήθηκαν ότι δεν μπορούν να προσφέρουν αξιόπιστη πληροφορία και μάλλον λειτουργούν ως θόρυβος.

rental_id	0
duration	0
bike_id	0
end_date	0
end_station_id	229639
end_station_name	0
start_date	0
start_station_id	229639
start_station_name	0
end_station_logical_terminal	24139562
start_station_logical_terminal	24139562
end_station_priority_id	24139562
dtype: int64	

id	0
installed	0
latitude	0
locked	0
longitude	0
name	0
bikes_count	0
docks_count	0
nbEmptyDocks	0
temporary	0
terminal_name	0
install_date	81
removal_date	782
dtype: int64	

Figure 6: Missing τιμές των αρχείων CSV του πρωταρχικού συνόλου δεδομένων

Επιπλέον της απομάκρυνσης των εγγραφών που εμπεριέχουν *ελλειπή δεδομένα*, η εξάλειψη των outliers κατέχει εξίσου σημαντικό ρόλο και επηρεάζει το προκύπτον αποτέλεσμα της διαδικασίας *εξόρυξης γνώσης*. Προς αυτή την κατεύθυνση, κρίθηκε αναγκαίο διαδρομές με συνολική διάρκεια μικρότερη του ενός λεπτού να μην ληφθούν υπόψη, ενώ συνολικά διατηρήθηκαν μόνο διαδρομές με διάρκεια εντός του διαστήματος $[Q1 - 1.5 \cdot IQR, Q3 + 1.5 \cdot IQR]$, όπου $IQR = Q3 - Q1$, γνωστή και ως *IQR (Interquartile Range) μέθοδος* για τον εντοπισμό outliers.

Όμοια με την απομάκρυνση outliers σε επίπεδο διαδρομών, αναζητούμε πιθανούς σταθμούς οι οποίοι ενδεχομένως να αποτελούν outliers. Για το σκοπό αυτό, ακολουθεί η απεικόνιση

[5]Γενικά μιλώντας, στην περίπτωση αριθμητικών χαρακτηριστικών υπάρχει μία πληθώρα από επιλογές, σχετικά με την διαχείριση *ελλιπών δεδομένων*, με την πιο απλή να είναι η αντικατάστασή τους με τον μέσο όρο ή τον ενδιάμεσο του εκάστοτε χαρακτηριστικού. Για κατηγορικά δεδομένα, όμως, η επιλογή για τον τρόπο διαχείρισης των *missing values* δεν είναι τόσο απλή.

όλων των διαθέσιμων σταθμών ενοικίασης σε έναν απλοϊκό χάρτη (βλ. *Figure 7*), με βάση τον οποίο διαπιστώνεται ότι, γεωγραφικά, κανένας σταθμός δεν μπορεί να θεωρηθεί με σιγουριά ως outlier. Στην σχετική βιβλιογραφία, συναντήσαμε ορισμένες φορές την θεώρηση σταθμών ως outliers με βάση την συνολική (χαμηλή) κίνηση, εισερχόμενη και εξερχόμενη, σε αυτούς. Παρόλο αυτά, στο πλαίσιο της συγκεκριμένης εργασίας, αποφασίστηκε να μην απομακρυνθούν έγκυροι σταθμοί, ακόμα και εάν αυτοί δεν είναι αρκετά δημοφιλείς.

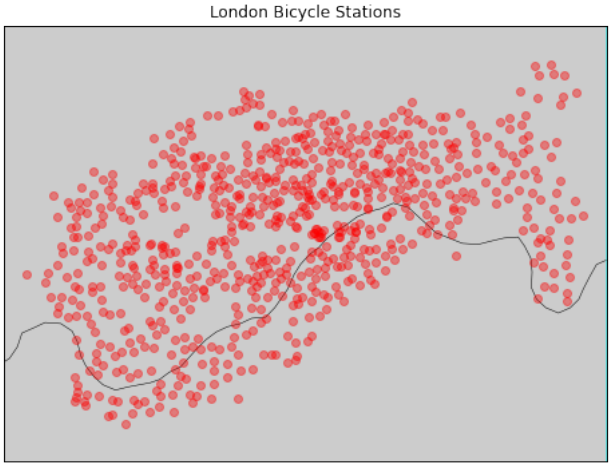


Figure 7: Σταθμοί ενοικίασης ποδηλάτων στο Λονδίνο

Έπειτα από την διατήρηση των κατάλληλων εγγραφών στα αρχεία διαδρομών και σταθμών ποδηλάτων, πραγματοποιείται συνένωση (*merge*) με τα πρόσθετα σύνολα δεδομένων που συλλέχθηκαν, έχοντας ως στόχο την δημιουργία συνδυασμένων στοχευμένων datasets, με βάση τα οποία θα καταστεί εφικτή η εξόρυξη όσο το δυνατό περισσότερης γνώσης, μετέπειτα. Προς αυτή την κατεύθυνση, κάθε σταθμός ενοικίασης αποκτά ζώνη (A, B, C), η οποία ταυτίζεται με την ζώνη στην οποία ανήκει ο πλησιέστερος σε αυτόν σταθμός μετρό, ενώ ταυτόχρονα η απόσταση από εκείνον χαρακτηρίζεται ως *Very Close*, *Close* και *Long*, σε μία προσπάθεια να εισαχθεί πληροφορία σχετικά με τον βαθμό που ο εκάστοτε σταθμός ενοικίασης πλαισιώνεται από το δίκτυο μετρό του Λονδίνου. Ομοίως, το dataset που εμπεριέχει τις διαδρομές συνενώνεται με εκείνο που φέρει τα δεδομένα καιρού, ώστε να είναι δυνατός ο προσδιορισμός των καιρικών συνθηκών, ανά διαδρομή. Τέλος, προκειμένου να έχουμε όλη την διαθέσιμη πληροφορία συγκεντρωμένη, πραγματοποιείται συνένωση των δύο εμπλουτισμένων datasets ενοικιάσεων και σταθμών ενοικίασης ποδηλάτων.

Για να μην αντιμετωπίσουμε πρόβλημα με την διαθέσιμη μνήμη RAM και τον χρόνο εκτέλεσης των προγραμμάτων μας, δεδομένου ότι τα εντοπιζόμενα patterns δεν μεταβάλλονται κατά πολύ από χρόνο σε χρόνο, αποφασίσαμε να εστιάσουμε αποκλειστικά στο έτος 2016, από εδώ και στο εξής.

2.2 Δημιουργία Βοηθητικών Χαρακτηριστικών (Feature Generation)

Αδιαμφισβήτητα, καθόλη την διαδικασία της *εξόρυξης γνώσης* από τα δεδομένα, η δημιουργία πρόσθετων χαρακτηριστικών διαδραματίζει πολύ σημαντικό ρόλο, τόσο κατά την εξερεύνηση των δεδομένων, μέσω για παράδειγμα διαφόρων ειδών οπτικοποιήσεων, όσο και κατά την φάση εφαρμογής τεχνικών μηχανικής μάθησης.

Από εδώ και στο εξής, στο πλαίσιο της συγκεκριμένης εργασίας, εστιάζουμε την προσοχή μας στο πρόβλημα ενοικίασης ποδηλάτων με δύο διαφορετικούς τρόπους, είτε σε επίπεδο διαδρομών[6], είτε σε επίπεδο σταθμών[7]. Πρακτικά, αυτές οι προσεγγίσεις αποτελούν δύο όψεις του ίδιου νομίσματος, με την διαφορά ότι μας επιτρέπουν να εξάγουμε πρόσθετη γνώση, ανάλογα με το που επικεντρώνεται το ενδιαφέρον μας κάθε φορά.

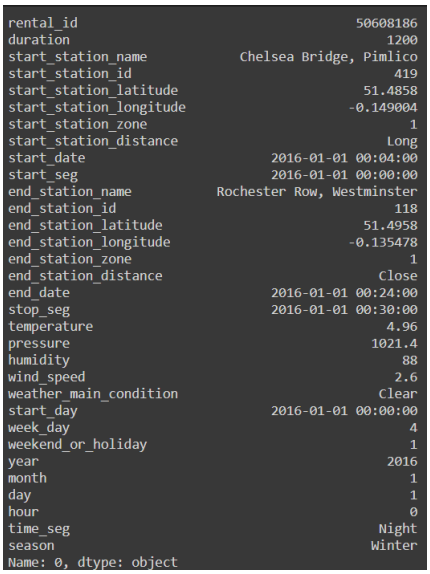
Ειδικότερα, όσον αφορά το κομμάτι των διαδρομών, μας απασχολεί η μελέτη της διάρκειας καθώς και το πλήθος των ενοικιάσεων που πραγματοποιούνται, ανά ώρα. Προς αυτή την κατεύθυνση, έχοντας ως στόχο την δημιουργία χρήσιμων οπτικοποιήσεων και την επιτέλεση ορισμένων απαραίτητων λειτουργιών ευκολότερα, κρίθηκε αναγκαία η *δημιουργία* των ακόλουθων *χαρακτηριστικών* (*feature generation*), με βάση την προϋπάρχουσα πληροφορία. Σε κάθε εγγραφή διαδρομής, προστίθενται τα χαρακτηριστικά *start_seg* και *stop_seg* που είναι *Timestamps* του πλησιέστερου μισάωρου στην ημερομηνία αναχώρησής και προορισμού, αντίστοιχα, της διαδρομής αυτής. Τα συγκεκριμένα χαρακτηριστικά συμβάλουν στην δημιουργία των *input* και *output flows* που μελετούνται εν συνεχεία. Επιπρόσθετα, σε κάθε εγγραφή, προστίθενται ορισμένα κατηγορικά χαρακτηριστικά, όπως η ώρα (*hour*), η ήμερα της εβδομάδας (*week_day*), η ημέρα (*day*), ο μήνας (*month*) της ημερομηνίας έναρξης της διαδρομής. Με βάση αυτά τα χαρακτηριστικά, προστίθεται επιπλέον το χαρακτηριστικό *weekend_or_holiday*, το οποίο υποδεικνύει εάν η ημερομηνία έναρξης αντιστοιχεί σε Σαββατοκύριακο ή κάποια επίσημα δηλωμένη αργία της Βρετανίας. Τέλος, δημιουργείται το κατηγορικό χαρακτηριστικό *season*, που υποδεικνύει την εποχή του χρόνου που εκτελέστηκε η διαδρομή[8]. Τέλος, δημιουργήθηκε το χαρακτηριστικό *weather_main_condition* το οποίο αποτελεί *Timestamp* με την ημερομηνία και ώρα έναρξης, στρογγυλοποιώντας προς τα κάτω στην πλησιέστερη ώρα, βοηθώντας μας αφενός να γίνει η συνένωση με τα δεδομένα καιρού (που ήταν διαθέσιμα ανά ώρα), καθώς και αφετέρου να γίνει η συγκέντρωση της μέσης διάρκειας διαδρομής και της μέσης ζήτησης ανά ώρα, ώστε να προκύψουν τα απαραίτητα visualizations. Η τελική μορφή του συνόλου δεδομένων που περιλαμβάνει τις διαδρομές ενοικίασης, παρουσιάζεται ακολούθως, απεικονίζοντας την πρώτη του εγγραφή (βλ. *Figure 8*).

Σχετικά με τους σταθμούς, χρησιμοποιώντας τα πρόσθετα χαρακτηριστικά που δημιουργήθηκαν ανά εγγραφή, τα οποία συζητήθηκαν αναλυτικά στην προηγούμενη παράγραφο, δημιουργήθηκε ένα επιπλέον dataset (*dat.csv*), όπου κάθε εγγραφή αντιστοιχεί σε ένα συγκεκριμένο σταθμό ενοικίασης, για κάποιο συγκεκριμένο *timeslot*, έχοντας διαχωρίσει τον χρόνο ανά μισάωρο. Έπειτα από την κατάλληλη επεξεργασία που απαιτήθηκε, παρουσιάζονται ακολούθως πέντε τυχαία επιλεγμένες εγγραφές του νεοσύστατου *dat.csv* (βλ. *Figure 9*).

[6]Σε αυτή την περίπτωση, μελετώνται οι διαδρομές μεμονωμένα ή στην καλύτερη περίπτωση ανά ώρα, όπως θα γίνει καλύτερα κατανοητό εν συνεχεία.

[7]Εδώ μας ενδιαφέρουν κυρίως οι σταθμοί, με βάση την εισερχόμενη και εξερχόμενη κίνηση σε αυτούς, ανά μισή ώρα.

[8]Με βάση το συγκεκριμένο χαρακτηριστικό θα προκύψουν ενδιαφέρουσες απεικονίσεις, που θα υποδηλώνουν την διαφορά στο μέσο demand και duration, από εποχή σε εποχή, ειδοποιός διαφορά των οποίων είναι στην ουσία οι καιρικές συνθήκες. Παρόλο αυτά αποφασίσαμε να αποφύγουμε απεικονίσεις με βάση τα χαρακτηριστικά καιρού, δεδομένου ότι είχαμε στην διάθεση μας δεδομένα καιρού ανά ώρα μόνο, εισάγοντας με αυτό τον τρόπο, άθελά μας, κάποιου είδους επιπλέον θόρυβο. Το συγκεκριμένο εμπόδιο, επομένως, ξεπερνιέται με την δημιουργία του κατηγορικού χαρακτηριστικού *season*, όπως προαναφέρθηκε, το οποίο πρακτικά αποτελεί κάποιου είδους *binning*, οδηγώντας μας προφανώς σε πιο ασφαλή συμπεράσματα.



rental_id	50608186
duration	1200
start_station_name	Chelsea Bridge, Pimlico
start_station_id	419
start_station_latitude	51.4858
start_station_longitude	-0.149004
start_station_zone	1
start_station_distance	Long
start_date	2016-01-01 00:04:00
start_seg	2016-01-01 00:00:00
end_station_name	Rochester Row, Westminster
end_station_id	118
end_station_latitude	51.4958
end_station_longitude	-0.135478
end_station_zone	1
end_station_distance	Close
end_date	2016-01-01 00:24:00
stop_seg	2016-01-01 00:30:00
temperature	4.96
pressure	1021.4
humidity	88
wind_speed	2.6
weather_main_condition	Clear
start_day	2016-01-01 00:00:00
week_day	4
weekend_or_holiday	1
year	2016
month	1
day	1
hour	0
time_seg	Night
season	Winter

Name: 0, dtype: object

Figure 8: Πρώτη εγγραφή του *bike_ hires.csv*, μετά την διαδικασία του *feature generation*

	station_id	time	in_flow_count	out_flow_count
8269461	340	2016-09-17 10:30:00	3.0	1.0
11214736	101	2016-05-12 08:00:00	15.0	9.0
11724152	519	2016-05-11 04:00:00	0.0	0.0
3655138	726	2016-01-21 17:00:00	0.0	1.0
717465	213	2016-11-03 04:30:00	0.0	0.0

Figure 9: Πέντε τυχαία επιλεγμένες εγγραφές του *dat.csv*

2.3 Εξερεύνηση Δεδομένων (Data Exploration)

Πλέον, έχοντας προηγηθεί η φάση της *προεπεξεργασίας* και *δημιουργίας* των απαιτούμενων *χαρακτηριστικών* και συνόλων δεδομένων, είμαστε έτοιμοι για την φάση *εξερεύνησης των δεδομένων*, η οποία ίσως αποτελεί μια εκ των πιο σημαντικών ενοτήτων της συγκεκριμένης εργασίας.

Αρχικά παρουσιάζονται (βλ. *Figure 10*) οι κατανομές των αριθμητικών χαρακτηριστικών, με χρήση *kernel density εκτιμητών*, οι οποίοι προσεγγίζουν τις πραγματικές κατανομές με βέλτιστο τρόπο, ειδικά όταν έχουμε δεδομένα μεγάλης κλίμακας, όπως συμβαίνει στην συγκεκριμένη περίπτωση. Με αυτό τον τρόπο αποκτούμε μία πρώτη εικόνα για τον τρόπο με τον οποίο κατανέμονται οι τιμές της θερμοκρασίας εδάφους, της ατμοσφαιρικής πίεσης, της υγρασίας, καθώς και της ωριαίας ζήτησης και διάρκειας διαδρομής. Έπειτα παρουσιάζονται τα αντίστοιχα *bar-charts* για τα κατηγορικά χαρακτηριστικά (βλ. *Figure 11*).

Έπειτα παρουσιάζονται ακολούθως τα *box-plots* της ζήτησης ποδηλάτων συνολικά, ανά εποχή, ανά ώρα της ημέρας και στην περίπτωση εργάσιμης-μη εργάσιμης ημέρας (βλ. *Figure 12*), με στόχο την ποιοτική κατανόηση της διασποράς του εκάστοτε χαρακτηριστικού προς μελέτη. Εύκολα παρατηρούμε ότι ο *median* και το *IQR* του demand είναι σχεδόν τα ίδια κατά την περίοδο του χειμώνα και της άνοιξης, ενώ αντίστοιχα είναι ίδια κατά την περίοδο του καλοκαιριού και του φθινοπώρου, λαμβάνοντας ελαφρώς υψηλότερες τιμές σε σχέση με την περίπτωση χειμώνα-άνοιξης. Σε επίπεδο ώρας εντός της ημέρας, παρατηρούμε σημαντική

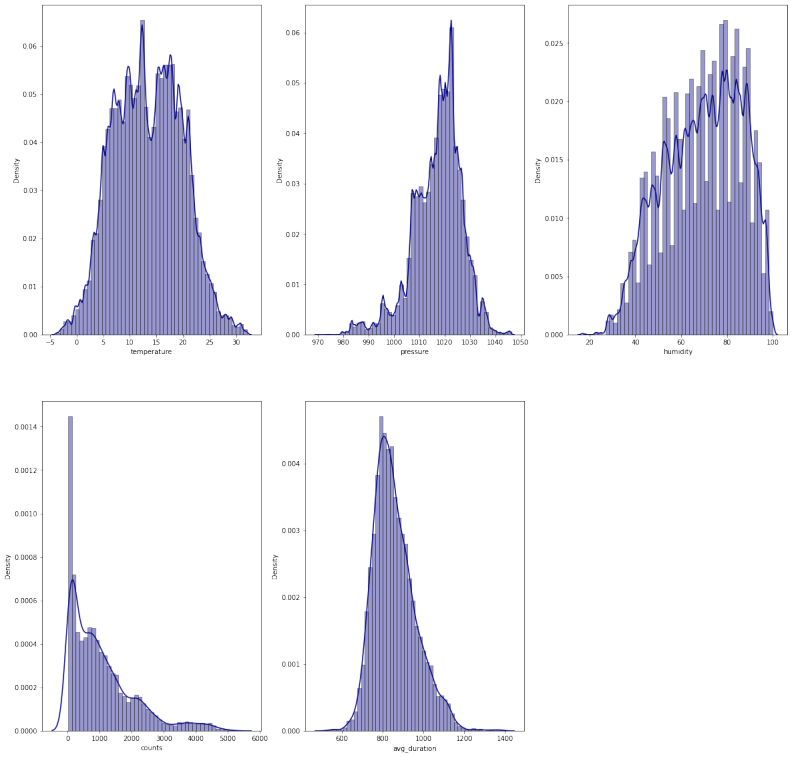


Figure 10: Κατανομές αριθμητικών δεδομένων

αύξηση του demand τις πρωινές ώρες (07.00 - 10.00), καθώς και τις απογευματινές ώρες (16.00 - 19.00), πράγμα που είναι αναμενόμενο. Τέλος σε επίπεδο εργάσιμης-μη εργάσιμης μέρας (Σαββατοκύριακο ή επίσημη αργία), παρατηρούμε ότι ο *median* και το *IQR* στην δεύτερη περίπτωση είναι αισθητά μικρότερο σε σχέση με την πρώτη περίπτωση, δεδομένου ότι δεν υπάρχει τόσο έντονη ζήτηση, κυρίως τις πρωινές και απογευματινές ώρες, όπως συμβαίνει τις εργάσιμες ημέρες, αντιπροσωπεύοντας μια πιο smooth καμπύλη.

Εν συνεχεία ακολουθεί ο πίνακας συσχέτισης των αριθμητικών δεδομένων (βλ. *Figure 13*), όπου για κάθε ζεύγος χαρακτηριστικών έχει υπολογιστεί ο *συντελεστής συσχέτισης Pearson*. Προφανώς το συγκεκριμένο μητρώο είναι συμμετρικό, με διαγώνια στοιχεία την μονάδα, πράγμα που δικαιολογεί το γεγονός ότι ακολούθως απεικονίζεται μόνο το κάτω τριγωνικό του μέρος. Παρατηρώντας προσεκτικά τον πίνακα συσχέτισης, εύκολα διαπιστώνουμε την σχετικά έντονη αρνητική συσχέτιση μεταξύ υγρασίας - θερμοκρασίας, ταχύτητας ανέμου - ατμοσφαιρικής πίεσης, ταχύτητας ανέμου - υγρασίας, καθώς και μεταξύ ζήτησης - υγρασίας και διάρκειας διαδρομής - υγρασίας. Αξίζει να σημειωθεί, σε αυτό το σημείο, ότι τα δύο τελευταία ευρήματα, σχετικά με την αισθητή αρνητική συσχέτιση μεταξύ ζήτησης και διάρκειας διαδρομής με την υγρασία, είναι πραγματικά μη αναμενόμενα. Στον αντίποδα, παρατηρείται σημαντική θετική συσχέτιση μεταξύ ζήτησης και διάρκειας διαδρομής με την θερμοκρασία εδάφους, πράγμα που είναι προφανώς αναμενόμενο εξ' αρχής.

Ακολούθως, παρουσιάζεται το *Cramer's V μητρώο* (βλ. *Figure 14*), το οποίο αποτελεί το ανάλογο του μητρώου συσχέτισης στην περίπτωση των κατηγορικών δεδομένων. Για κάθε ζεύγος κατηγορικών χαρακτηριστικών, υπολογίζεται ο *Cramer's V συντελεστής*, ο οποίος βασίζεται στον χ^2-test, όπως είχαμε συζητήσει και στο μάθημα. Με βάση το *Cramer's V μητρώο*, επομένως, διαπιστώνουμε ότι τα χαρακτηριστικά *month* - *season* και

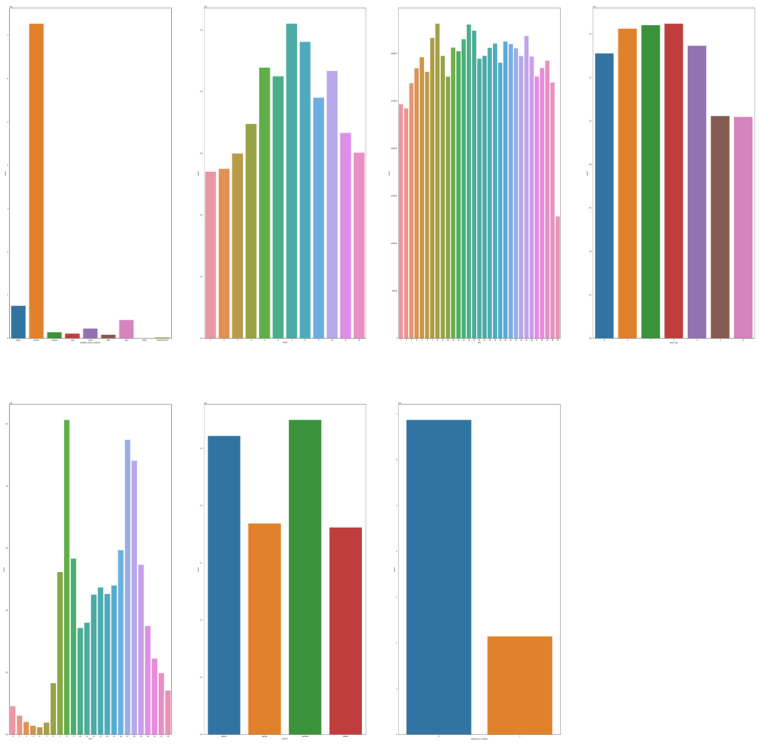


Figure 11: Κατανομές Κατηγορικών δεδομένων

weekend_or_holiday - *week_day* συσχετίζονται ισχυρά μεταξύ τους. Αυτό είναι απόλυτα λογικό, μιας και στην πρώτη περίπτωση το κατηγορικό χαρακτηριστικό *season* προέκυψε με βάση το χαρακτηριστικό *month*. Ομοίως, ανάλογα με το *week_day* καθορίζεται εάν η διαδρομή εκτελείται εντός του Σαββατοκύριακου, πλην ορισμένων ελάχιστων εξαιρέσεων όπου συμπεριλαμβάνονται επίσημες αργίες.

Εν συνεχεία έπονται ορισμένες χρήσιμες απεικονίσεις, με βάση τις οποίες θα εξαχθεί σημαντική πληροφορία που υποκρύπτεται στα δεδομένα μας. Αρχικά παρουσιάζεται η μέση μηνιαία ζήτηση και διάρκεια διαδρομής (βλ. *Figure 15*). Σχετικά με την ζήτηση, παρατηρείται ραγδαία αύξηση κατά τους θερινούς και τους πρώτους μήνες του φθινοπώρου (Μάιος - Οκτώβριος), σε σχέση με το υπόλοιπο έτος, πράγμα που είναι αναμενόμενο λόγω των καλύτερων καιρικών συνθηκών που επικρατούν στο Λονδίνο. Αντιθέτως, η μηνιαία διάρκεια διαδρομής δεν παρουσιάζει τόσο μεγάλη μεταβολή από μήνα σε μήνα, πρακτικά είναι σχεδόν ομοιόμορφη, απλώς εντοπίζεται κάποια πολύ ελαφριά αύξηση κατά τους μήνες του καλοκαιριού, παρομοίως.

Ακολούθως απεικονίζεται η μέση ωριαία ζήτηση και διάρκεια, ανάλογα με την εποχή (βλ. *Figure 16*). Αρχικά, οφείλουμε να επισημάνουμε ότι τα patterns είναι όμοια και στις δύο περιπτώσεις, ανεξαρτήτως εποχής. Ειδικότερα, όσον αφορά την μέση ζήτηση παρατηρούμε ότι κυμαίνεται σε παρόμοια επίπεδα χειμώνα - άνοιξη και καλοκαίρι - φθινόπωρο. Έντονες διαφορές στην ζήτηση, μεταξύ διαφορετικών εποχών, παρατηρούνται μετά τις πρώτες πρωινές ώρες (8.00 - 9.00+). Όπως είχε επισημανθεί προηγουμένως, κατά την μελέτη του σχετικού *box-plot*, συστηματικά παρουσιάζεται ραγδαία αύξηση της ζήτησης 7.00 - 10.00 και 16.00 - 19.00, γεγονός που είναι ως ένα βαθμό αναμενόμενο. Επιπλέον, αξίζει να σημειωθεί ότι το δεύτερο (απογευματινό) peak είναι πολύ πιο έντονο κατά την περίοδο του καλοκαιριού και του φθινοπώρου, όπου οι καιρικές συνθήκες είναι ηπιότερες καθώς νυχτώνει. Σχετικά

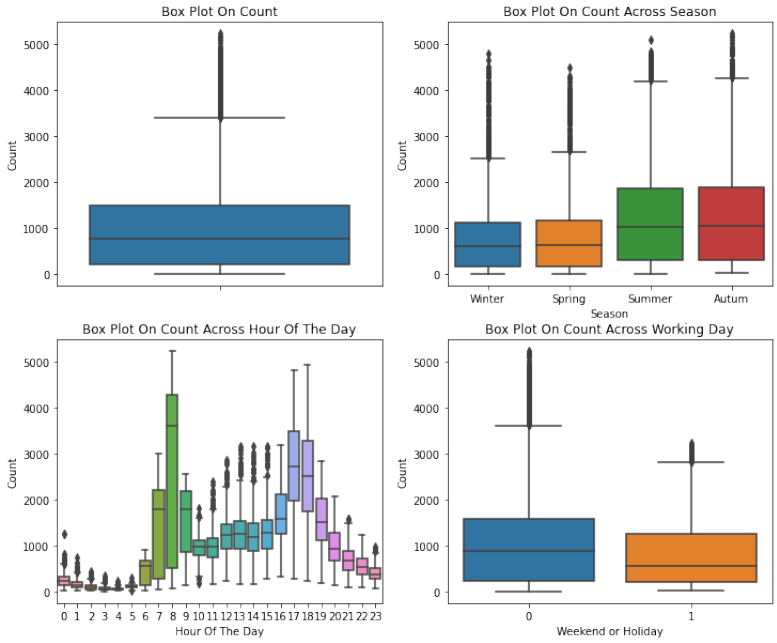


Figure 12: Ανάλυση outliers με βάση την ωριαία ζήτηση

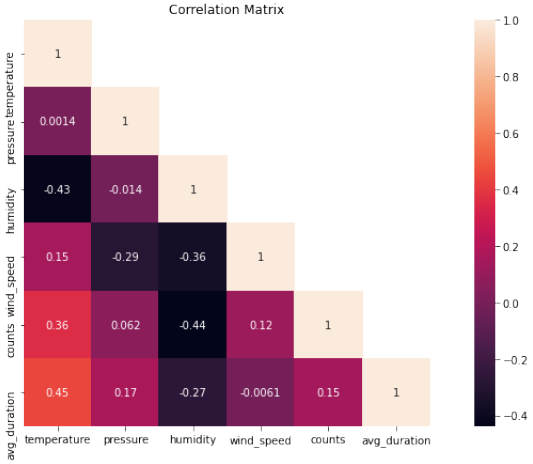


Figure 13: Πίνακας συσχέτισης αριθμητικών χαρακτηριστικών

με την μέση διάρκεια, λαμβάνοντας υπόψη μας ότι ο χρόνος μετριέται σε δευτερόλεπτα, οδηγούμαστε στο συμπέρασμα ότι δεν μεταβάλλεται σημαντικά μεταξύ των διαφορετικών ωρών της ημέρας. Κατά κύριο λόγο εντοπίζονται δύο peaks τις βραδινές (00.00 - 05.00) και μεσημεριανές (14.00 - 18.00) ώρες, χωρίς όμως στην πραγματικότητα να υπάρχουν τόσο έντονες μεταβολές.

Έπειτα γίνονται τα αντίστοιχα διαγράμματα για τις εργάσιμες - μη εργάσιμες ημέρες, ανά ώρα του 24ωρου (βλ. *Figure 17*). Γενικά μιλώντας, εστιάζοντας στις ημέρες ενός έτους, ένας λογικός διαχωρισμός αυτών είναι σε εργάσιμες και τις μη εργάσιμες ημέρες (Σαββατοκύριακα ή επίσημες αργίες). Υπό αυτό το πρίσμα, τα διαγράμματα του *Figure 17* μπορούν να θεωρηθούν ως μία αποσύνθεση (*decomposition*) των αντίστοιχων διαγραμμάτων του *Figure 16*, στις δύο αυτές συνιστώσες. Ως άμεσο αποτέλεσμα, σχετικά με την μέση ζήτηση ανά ώρα, παρατηρούμε ότι στην περίπτωση των μη εργάσιμων ημερών, παρουσιάζεται ένα μόνο peak με μεγαλύτερο εύρος (σχεδόν τα 2/3 του 24ωρου), έναντι δύο στενότερων peaks κατά

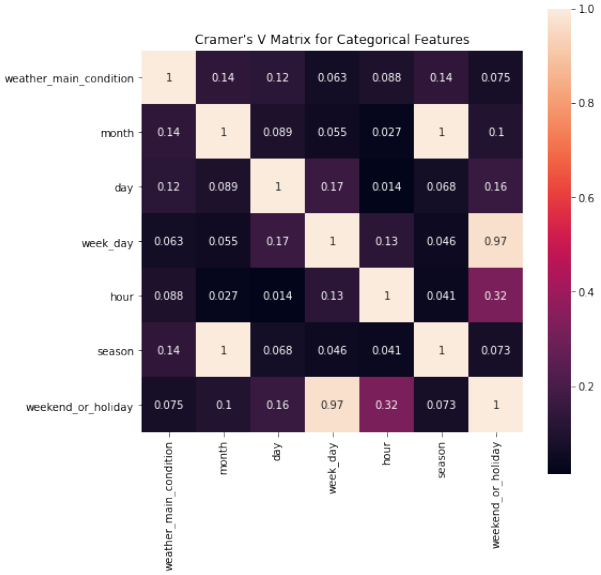


Figure 14: *Cramer's V μητρώο* για κατηγορικά δεδομένα

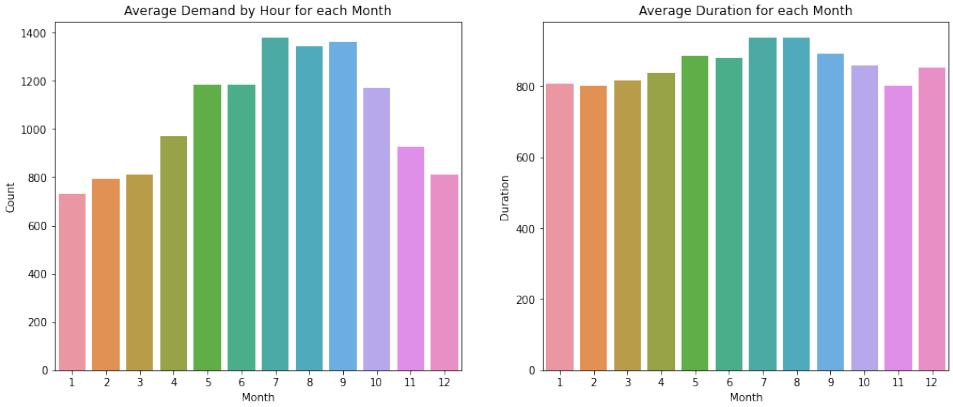


Figure 15: Μέση ζήτηση ενοικίασης και μέση διάρκεια διαδρομής ανά μήνα

τις πρωινές και απογευματινές ώρες, όταν η πλειοψηφία του κόσμου πηγαίνει και γυρνάει από τον χώρο εργασίας τους, αντίστοιχα. Τέλος, όσον αφορά την μέση διάρκεια διαδρομής, σε σχέση με τις εργάσιμες ημέρες, παρατηρείται αυξημένη μερικά λεπτά, κατά το μεγαλύτερο κομμάτι του 24ωρού.

Εν συνεχεία παρουσιάζονται το ανάλογο του *Figure 15*, με βάση εάν ο σταθμός αφετηρίας είχε "βαπτιστεί" στην ζώνη A, B ή C, χρησιμοποιώντας *stack-bar charts*. Αξίζει να υπενθυμίσουμε σε αυτό το σημείο, ότι κάθε σταθμός ενοικίασης ποδηλάτων "βαπτίστηκε" στην αντίστοιχη ζώνη που άνηκε ο πλησιέστερος σε αυτόν σταθμός μετρό, επιδιώκοντας με αυτό τον τρόπο εισάγουμε στο παιχνίδι την εξάρτηση από την απόσταση του σταθμού αφετηρίας/προορισμού από το κέντρο της πόλης. Σχετικά με την μέση ζήτηση, εύκολα γίνεται αντιληπτό ότι οι περισσότερες ενοικιάσεις πραγματοποιούνται από σταθμούς αφετηρίας που ανήκουν στην ζώνη A, στην οποία περιλαμβάνεται το κεντρικότερο κομμάτι της πόλης, με πληθώρα τουριστών καθημερινά, αλλά και εντονότατη οικονομική δραστηριότητα. Έπειτα έπεται η ζώνη B, ενώ παρατηρούμε ελάχιστη ζήτηση σε σταθμούς της ζώνης C, που απέχουν κατά κανόνα αρκετά από τα κεντρικά σημεία της πόλης. όσον αφορά την διάρκεια διαδρομής, αισθητά μεγαλύτερη είναι σε περιπτώσεις που ο σταθμός αφετηρίας ήταν στην ζώνη C, σε

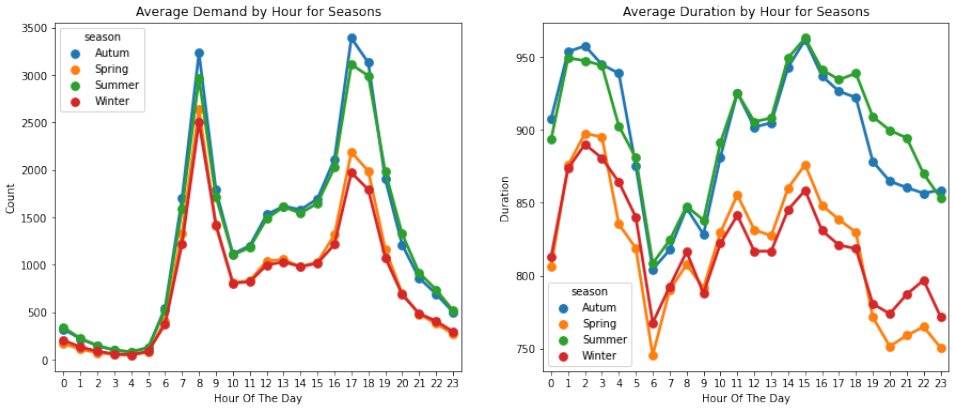


Figure 16: Μέση ζήτηση ενοικίασης και μέση διάρκεια διαδρομής ανά ώρα, ανάλογα με την εποχή

Figure 17: Μέση ζήτηση ενοικίασης και μέση διάρκεια διαδρομής ανά ώρα, ανάλογα με το εάν είναι εργάσιμη ημέρα ή όχι

σχέση με τις ζώνες A, B, πράγμα που είναι απόλυτα λογικό και αναμενόμενο. Σχεδόν ίδια είναι η εικόνα και στην περίπτωση των σταθμών προορισμού, τόσο σχετικά με την ζήτηση, όσο και με την διάρκεια διαδρομής[9].

Ακολούθως παρουσιάζεται μέση ζήτηση και η μέση διάρκεια διαδρομών που έχουν ως αφετηριά σταθμούς που απέχουν "Very Close", "Close" και "Long"[10] από τον πλησιέστερο, σε αυτούς, σταθμό μετρό (βλ. *Figure 19*). Αρχικά παρατηρούμε ότι η διάρκεια διαδρομής φαίνεται να είναι ίδια και στις τρεις αυτές περιπτώσεις χωρίς να μας προσφέρει κάποια χρήσιμη πληροφορία. Αντιθέτως, στην περίπτωση της μέση ζήτησης γίνεται εμφανές ότι η μέση ζήτηση σε σταθμούς που βρίσκονται πολύ κοντά (*Very Close*) και σχετικά κοντά (*Close*, $d < 700m$) σε κάποιον σταθμό μετρό είναι παρόμοια, ανεξαρτήτως μήνα. Δεν συμβαίνει όμως το ίδιο και με σταθμούς που απέχουν αρκετά (*Long*) από κάποιο σταθμό μετρό, πράγμα που ακούγεται λογικό. Ομοίως με το *Figure 18*, η το αντίστοιχο διάγραμμα σχετικά με τους σταθμούς προορισμού είναι σχεδόν το ίδιο και αυτό το λόγο κρίνεται αναγκαίο να παραληφθεί.

[9] Το αντίστοιχο διάγραμμα παρουσιάζεται στο επισυναπτόμενο Python Notebook, το οποίο εμπεριέχει όλον τον απαιτούμενο κώδικα Python της συγκεκριμένης εργασίας, και παραλείπεται για εξοικονόμηση χώρου της παρούσας αναφοράς.

[10] Παρατίθεται σε αυτό το σημείο η αγγλική ορολογία, μιας και αυτές είναι μόνο οι δυνατές τιμές για το συγκεκριμένο κατηγορικό χαρακτηριστικό.

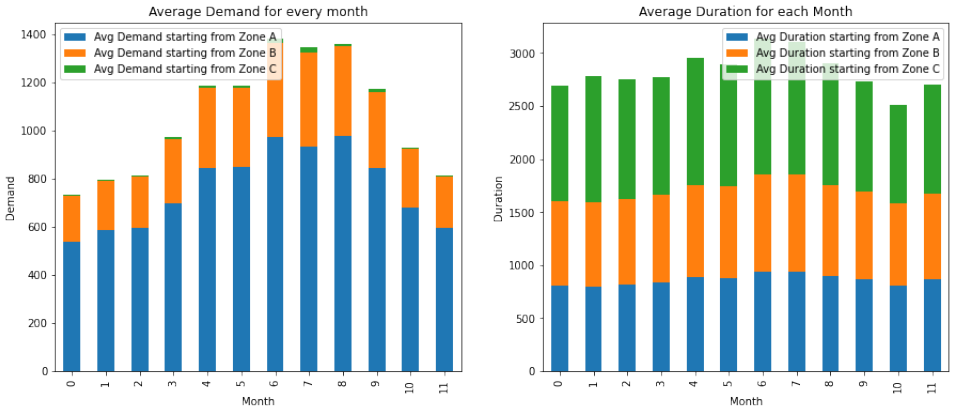


Figure 18: Μέση ζήτηση ενοικίασης και μέση διάρκεια διαδρομής, ανά μήνα, ανάλογα με την ζώνη του σταθμού αφετηρίας

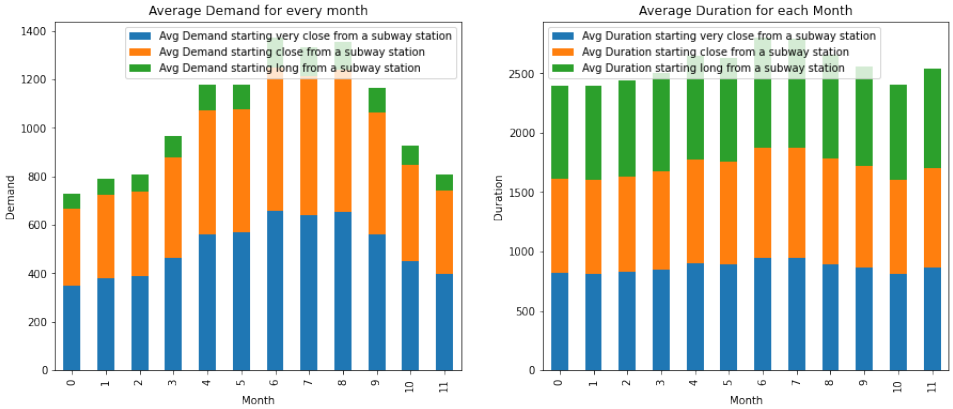


Figure 19: Μέση ζήτηση ενοικίασης και μέση διάρκεια διαδρομής, ανά μήνα, ανάλογα με τό πόσο κοντά βρίσκεται ο σταθμός αφετηρίας στον πλησιέστερο σταθμό μετρό

Όσον αφορά σε επίπεδο σταθμών τώρα, αξιοποιώντας τα πρόσθετα datasets που δημιουργήσαμε σχετικά με τα *input, output flows*, προνοώντας να είναι στην κατάλληλη μορφή που απαιτούνταν για ευκολότερη επεξεργασία, είμαστε σε θέση να παρουσιάσουμε ορισμένες χρήσιμες οπτικοποιήσεις. Σε πρώτη φάση, παρατηρώντας τον τρόπο με τον οποίο κατατάσσονται οι σταθμοί (βλ. *Figure 7*), εύκολα διαπιστώνουμε τον σχεδόν ομοιόμορφο διαμοιρασμό τους στον χώρο. Παρόλο αυτά, όπως είναι φυσικό, σε κεντρικές περιοχές του Λονδίνου αναμένεται να υπάρχει μεγαλύτερη πυκνότητα σε σταθμούς ενοικίασης, κάτι που προκύπτει ξεκάθαρα και από το *Figure 7*.

Εν συνεχεία, εκμεταλλευόμενοι τα *δεδομένα ροών* (*flows data*) που σχηματίσαμε, καταλήγουμε ότι οι δέκα πιο συχνές διαδρομές, κατά τις εργάσιμες ημέρες, είναι οι εξής: (191, 191), (785, 785), (303, 303), (307, 307), (248, 248), (789, 789), (183, 74), (307, 404), (671, 729), (71, 154), όπου κάθε ζεύγος απαρτίζεται από τα ids των σταθμών αφετηρίας - προορισμού. Ομοίως, οι δέκα πιο συχνές διαδρομές τα Σαββατοκύριακα προέκυψε ότι είναι: (191, 191), (785, 785), (307, 307), (303, 303), (789, 789), (248, 248), (786, 785), (307, 191), (191, 248), (307, 303). Ακολούθως απεικονίζονται οι σταθμοί, στις δύο αυτές περιπτώσεις (βλ. *Figure 20*), ενώ έπεται ο απαραίτητος σχολιασμός.
Με βάση τους απλοϊκούς χάρτες που παρουσιάζονται παραπάνω, διαπιστώνουμε ότι οι δέκα πιο συχνά εμφανιζόμενες διαδρομές τις καθημερινές αφορούν σταθμούς στο κεντρικό και

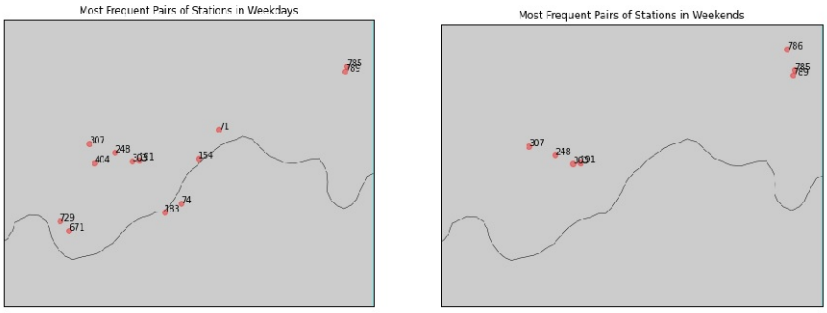


Figure 20: Ζεύγη σταθμών που είναι πιο δημοφιλή τις καθημερινές και τα Σαββατοκύριακα/αργίες, αντίστοιχα

βόριο ανατολικό Λονδίνο, όπου στην πλειοψηφία των διαδρομών ο σταθμός αφετηρίας και προορισμού ταυτίζονται[11]. Οι υπόλοιπες συχνότερες διαδρομές που εντοπίστηκαν, με διαφορετικό σταθμό αφετηρίας και προορισμού, προκύπτει ότι είναι πολύ σύντομες, ταξιδεύοντας σε κάποιον πολύ κοντινό σταθμό από το σημείο εκκίνησης, Στην περίπτωση των Σαββατοκύριακων, η κατάσταση παραμένει η ίδια λίγο πολύ, με την συμμετοχή φυσικά λιγότερων σε πλήθος σταθμών, εξαιτίας της χαμηλότερης ζήτησης. Αξίζει να σημειωθεί, ότι πληροφορία τέτοιου είδους, σχετικά με την δημοφιλία των σταθμών, είναι ιδιαίτερα χρήσιμη για την αποτελεσματικότερη οργάνωση των *BSS* και την καλύτερη κατανομή ποδηλάτων, καθώς επίσης και για τα κέρδη των εταιριών που δραστηριοποιούνται στον χώρο. Αυτός ήταν και ο βασικός λόγος που μας προέτρεψε να μην σταθούμε αποκλειστικά στα χαρακτηριστικά διαδρομής (ζήτηση, διάρκεια), αλλά να προσπαθήσουμε να εξαγάγουμε γνώση και για τους σταθμούς ενοικίασης αυτούς καθαυτούς, με βάση τις διαδρομές που πραγματοποιήθηκαν εντός του 2016.

Αξιοποιώντας και πάλι *flows datasets*, παρουσιάζονται ακολούθως οι δέκα πιο δημοφιλείς σταθμοί, δηλαδή εκείνοι με την μεγαλύτερη μέση ροή συνολικά (κόκκινο χρώμα). Επιπρόσθετα σκεφτήκαμε ότι μία ενδιαφέρουσα εναλλακτική ταξινόμηση, μεταξύ των σταθμών, θα προέκυπτε χρησιμοποιώντας τον αλγόριθμο *Google's PageRank*[12]. Στην περίπτωση αυτή, μεταχειριζόμαστε τους σταθμούς ως κορυφές ενός υποτιθέμενου γραφήματος, του οποίου οι ακμές έχουν ως βάρος το πλήθος των διαδρομών από τον ένα σταθμό στον άλλον. Με βάση την συγκεκριμένη μέθοδο, επιβραβεύονται με υψηλότερο score σταθμοί που συνδέονται με άλλους "σημαντικούς" σταθμούς, δημιουργώντας ένα δίκτυο αρκετά σημαντικών σταθμών. Όπως φαίνεται και ακολούθως (βλ. *Figure 21*), οι σταθμοί που προκύπτουν δεν είναι απαραίτητα και οι πιο δημοφιλής (δηλαδή εκείνοι με υψηλή συνολική ροή), αλλά αυτοί που διαθέτουν τις περισσότερες διαδρομές προς πολύ "σημαντικούς" σταθμούς. Αξίζει να σημειωθεί ότι, όπως αναφέρεται και στο [9], υπάρχουν ορισμένες "παθογένειες" σε τέτοιου είδους γραφήματα που πηγάζουν από προβλήματα του πραγματικού κόσμου, οι οποίες όμως

[11] Αυτός ακριβώς είναι και ο λόγος που επικεντρωθήκαμε εξ' αρχής στην ζήτηση και διάρκεια διαδρομής αποκλειστικά, χωρίς να λαμβάνουμε υπόψη την απόσταση διαδρομής, η οποία εμφανίζεται ιδιαίτερα συχνά στην βιβλιογραφία σε datasets που εμπεριέχουν διαδρομές ταξί. Εκεί πράγματι, είναι σχεδόν απίθανο το σημείο έναρξης και τερματισμού της κούρσας να ταυτίζονται. Στην περίπτωση των *BSS* τα πράγματα είναι διαφορετικά όμως, με αποτέλεσμα η γέννηση ενός χαρακτηριστικού σχετικού με την απόσταση διαδρομής, δεν αναμένεται να είναι αρκετά *informative*.

[12] https://en.wikipedia.org/wiki/PageRank

ξεπερνιόνται σχετικά απλά[13]. Παρόλο αυτά, η συγκεκριμένη μέθοδος είναι αρκετά δημοφιλής, ενώ διαθέτει μία μεγάλη γκάμα εφαρμογών. Με βάση το *Figure 21*, γίνεται εύκολα αντιλη-

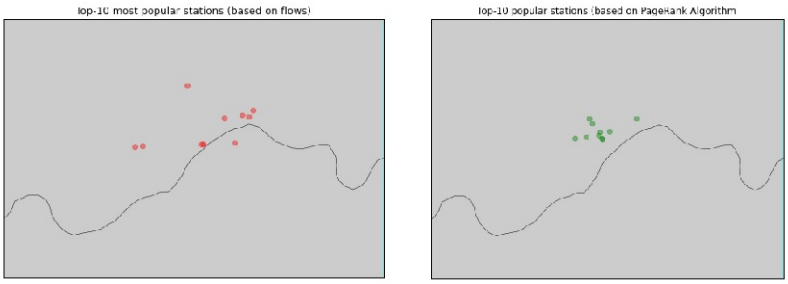


Figure 21: Πιο σημαντικοί σταθμοί με βάση της ροές (κόκκινο χρώμα) και με βάση το αποτέλεσμα που προκύπτει από την εφαρμογή της *PageRank μεθόδου* (πράσινο χρώμα)

πτό ότι το αποτέλεσμα των *top-10* σταθμών με βάση της υψηλότερες ροές κίνησης διαφέρει ελαφρώς από το αποτέλεσμα του *PageRank αλγορίθμου*. Στην τελευταία περίπτωση έχουν εντοπιστεί πιο κεντρικοί σταθμοί, έναντι της πρώτης, οι οποίοι μπορεί να μην διαθέτουν όλοι από αυτούς τις υψηλότερες ροές κινήσεις, όμως η μέθοδος τους εντοπίζει ως ιδιαίτερα "σημαντικούς" σταθμούς. Διαισθητικά, το αποτέλεσμα του *PageRank αλγορίθμου* φαίνεται πιο πειστικό στην πράξη και σίγουρα πρέπει να ληφθεί υπόψιν[14].

Τέλος, στους ακόλουθους απλοϊκούς χάρτες παρουσιάζονται οπτικά, ανά εποχή, οι σταθμοί με βάση το συνολικό τους *flow*, δίνοντας έμφαση σε εκείνους των οποίων η ροή ήταν πάνω από ένα προκαθορισμένο κατώφλι[15]. Εύκολα γίνεται αντιληπτό, ότι υψηλότερες ροές παρατηρούνται κατά κύριο λόγο σε κεντρικούς σταθμούς, πράγμα που είχε επισημανθεί και προηγουμένως καθώς μελετούσαμε την ζήτηση με βάση την ζώνη του εκάστοτε σταθμού. Κάτι τέτοιο, όμως, είναι απολύτως αναμενόμενο, αφού σε αυτές τις περιοχές εντοπίζεται μεγάλη οικονομική δραστηριότητα, ενώ επίσης υπάρχουν πληθώρα από αξιοθέατα, όπως είναι για παράδειγμα το *Βρετανικό Μουσείο*, τα οποία προσελκύουν έντονα το ενδιαφέρον των τουριστών κάθε χρόνο, παράγοντες που σίγουρα διαδραματίζουν σημαντικό ρόλο στο παρόν προκύπτον αποτέλεσμα. Τέλος, με βάση το συγκεκριμένο *Figure*, παρατηρούμε ότι οι σταθμοί με τα περισσότερα *flows* αλλάζουν από εποχή σε εποχή, ενώ επιπλέον οι ροές είναι προφανώς εντονότερες κατά το καλοκαίρι και το φθινόπωρο, καθώς φτάνουν στο ναδίρ κατά την περίοδο των χειμερινών μηνών.

Σε αυτό το σημείο ολοκληρώνεται η φάση *εξερεύνησης των δεδομένων*, κατά την οποία επικεντρωθήκαμε τόσο σε επίπεδο διαδρομής (ζήτηση, διάρκεια) όσο και σε επίπεδο σταθμών (δημοφιλία σταθμών, ροές κίνησης). Ακολούθως έπεται η εφαρμογή τεχνικών Μηχανικής Μάθησης (*Κεφάλαιο 3*), καθώς και τέλος η σύνοψη των συμπερασμάτων που καταφέραμε να συλλέξουμε σε συνδυασμό με ορισμένες μελλοντικές κατευθύνσεις που προτείνουμε (*Κεφάλαιο 5*).

[13]Στο [9] πραγματοποιείται επίσης εκτενής συζήτηση για το πως η συγκεκριμένη μέθοδος μπορεί να γίνει *scale-up* για δίκτυα πολύ μεγάλης κλίμακας, πράγμα που θεωρούμε ιδιαίτερα ενδιαφέρον και άξιο προς επισήμανση.

[14]Τα αποτελέσματα των σταθμών που προκύπτουν στις δύο περιπτώσεις που συζητήθηκαν προηγουμένως, παρουσιάζονται σε μορφή *DataFrame* στο Python Notebook, το οποίο περιλαμβάνει όλο τον κώδικα που αναπτύχθηκε στο πλαίσιο της συγκεκριμένης εργασίας.

[15]Το κατώφλι αυτό έχει οριστεί σε τουλάχιστον 2000 εισερχόμενες και εξερχόμενες διαδρομές αθροιστικά, ανά εποχή, ενώ η ακτίνα τους είναι ανάλογη των *flows* τους, παρέχοντας το ανάλογο *visualization*

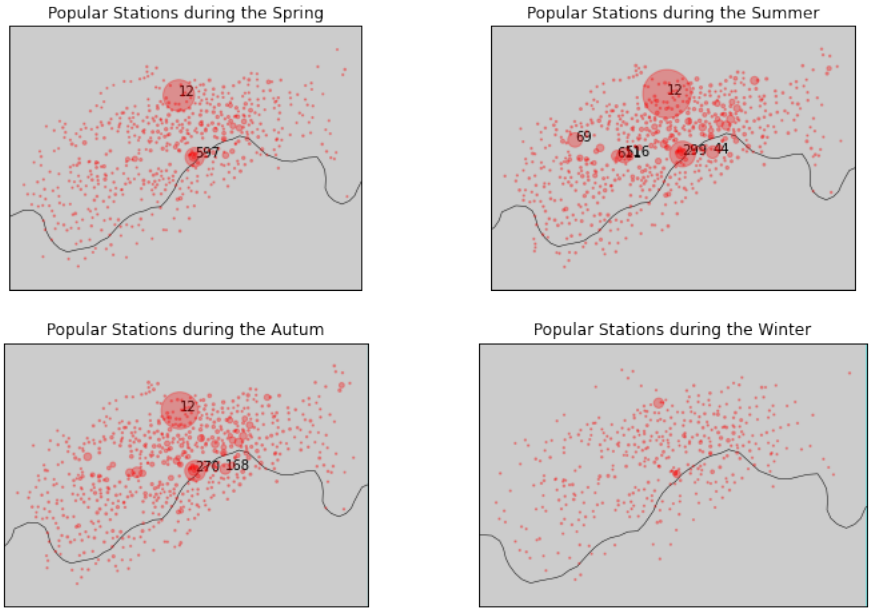


Figure 22: *Flows* σταθμών ανάλογα με την εποχή

3 Τεχνικές Μηχανικής Μάθησης

Στο παρόν κεφάλαιο πραγματοποιείται εφαρμογή μεθόδων *μηχανικής μάθησης*, με στόχο την περαιτέρω εξαγωγή γνώσης. Προς αυτή την κατεύθυνση επιλέξαμε να επικεντρωθούμε στην *πρόβλεψη ωριαίας ζήτησης* ποδηλάτων (*hourly demand prediction)*, καθώς και στην *συσταδοποίηση* (*clustering*) των σταθμών σε ομάδες, με βάση την τοποθεσία τους (*spatial clustering)*. Αυτή η προσέγγιση αποτελεί συνέχεια εκείνης που είχε επιλεγεί στο πλαίσιο του *Κεφαλαίου 2*, δηλαδή την εστίαση τόσο σε επίπεδο διαδρομής, όσο και σε επίπεδο σταθμών, αντίστοιχα.

3.1 Πρόβλεψη (Prediction)

Όπως προαναφέρθηκε, στην συγκεκριμένη υποενότητα σκοπός μας είναι η *πρόβλεψη της ωριαίας ζήτησης* ποδηλάτων. Αρχικά δοκιμάζουμε την επίδοση ενός *Random Forest Regressor*, που αποτελεί μία εκ των κλασικών *τεχνικών μηχανικής μάθησης*, δίνοντας μας ταυτόχρονα τροφή για συζήτηση και συγκρίσεις με τα μετέπειτα μοντέλα *βαθιάς μηχανικής μάθησης* που ακολουθούν, ονομαστικά το *GRU*, *LSTM*, *Bidirectional LSTM*), τα οποία είναι ίσως από τα πιο δημοφιλή σε προβλήματα *πρόβλεψης χρονοσειρών*, μιας και επιτυγχάνουν συνήθως πολύ καλές επιδόσεις[16]. Ακολούθως παρουσιάζεται το διάγραμμα της ζήτησης για τα έτη 2015 και 2016 (βλ. *Figure 23*). Εύκολα γίνεται αντιληπτή η ύπαρξη κάποιου είδους περιοδικότητας στην ωριαία ζήτηση ποδηλάτων. Με βάση την συγκεκριμένη παρατήρηση, κατ' εξαίρεση, έχοντας ως στόχο την βελτίωση της ποιότητας των αποτελεσμάτων που προκύπτουν από τις προαναφερθείσες *τεχνικές μηχανικής μάθησης*, χρησιμοποιούμε και τα δεδομένα του 2015.

[16]Στο επισυναπτόμενο Python Notebook περιέχεται και η περίπτωση επιτέλεσης *prediction* χρησιμοποιώντας ένα κλασικό *feed-forward νευρωνικό δίκτυο*, το οποίο αναμενόμενα δεν επιτυγχάνει καθόλου καλές επιδόσεις. Για αυτόν ακριβώς το λόγο, κρίναμε ότι δεν χρειάζεται να συμπεριληφθεί στην παρούσα αναφορά.

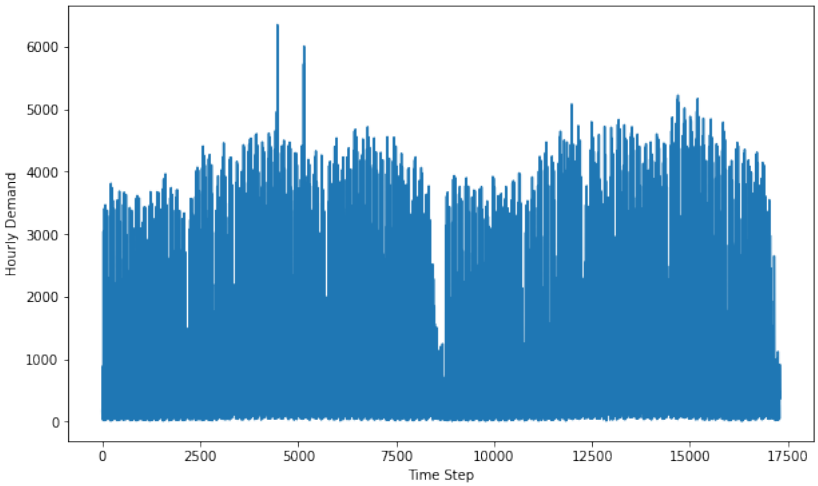


Figure 23: Μέση ωριαία ζήτηση ποδηλάτων για τα έτη 2015 και 2016

3.1.1 Random Forest (RF) Regressor

Γενικά μιλώντας, ο *Random Forest* αλγόριθμος μάθησης αποτελεί μία από τις κλασικότερες και πιο δημοφιλής μεθόδους της συγκεκριμένης περιοχής. Πρόκειται για μια *bagging* τεχνική, η οποία συνδυάζει καταλλήλως τις αποφάσεις ενός πλήθους από *δέντρα απόφασης* (*decission trees*), με στόχο την επίτευξη καλύτερου αποτελέσματος. Αντίστοιχη είναι και η ιδέα του *RF Regressor*, η οποία απεικονίζεται και ακολούθως (βλ. *Figure 24*), χρησιμοποιώντας 600 *δέντα απόφασης*, σαν *regressors βάσης*, στο συγκεκριμένο παράδειγμα.

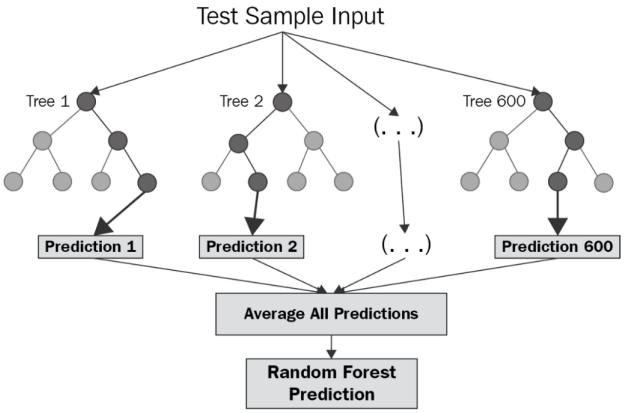


Figure 24: *Random Forest Regressor* [3]

Όσον αφορά το πρόβλημα που μελετούμε, σχετικά με *πρόβλεψη της ωριαίας ζήτησης*, έπειτα από αρκετές δοκιμές, επιλέξαμε να συμπεριλάβουμε τα (αριθμητικά) καιρικά χαρακτηριστικά (θερμοκρασία εδάφους, ατμοσφαιρική πίεση, υγρασία, ταχύτητα ανέμου), καθώς και τα κατηγορικά χαρακτηριστικά σχετικά με την εποχή, την γενική περιγραφή του καιρού, την χρονική περίοδο εντός της ημέρας και την ένδειξη σχετικά με το εάν κάποια μέρα είναι εργάσιμη ή όχι. Με βάση την εκτενή συζήτηση που έγινε στο *Κεφάλαιο 2*, γνωρίζουμε ότι τα συγκεκριμένα χαρακτηριστικά προσφέρουν σημαντική πληροφορία, ικανή να διαχωρίσει trends μεταξύ του, η οποία ταυτόχρονα είναι σχετικά abstract, βοηθώντας τα μοντέλα μας να μην οδηγηθούν σε overfitting. Προκειμένου να έχουμε μία κοινή βάση αναφοράς, τα ίδια ακριβώς

χαρακτηριστικά ακολουθούν και στα υπόλοιπα μοντέλα που έπονται.

Ακολούθως περιγράφεται εν συντομία η διαδικασία που ακολουθήσαμε. Αρχικά δημιουργούνται *dumy μεταβλητές* για τα κατηγορικά χαρακτηριστικά, ενώ τα αριθμητικά χαρακτηριστικά (δεδομένα καιρού και ωριαία ζήτηση) γίνονται normalized, αφαιρώντας από τις τιμές του εκάστοτε χαρακτηριστικού την μέση τιμή αυτού και διαιρώντας με την διασπορά του. Έπειτα τα αρχικά δεδομένα διασπώνται με τυχαίο τρόπο σε *training* kai *test sets*, σε ποσοστό $90\% - 10\%$, αντίστοιχα[17]. Χρησιμοποιώντας ως *regressors βάσης* 100 *δέντρα απόφασης*, πραγματοποιείται το *fit* του μοντέλου στα *training δεδομένα*, επιτυγχάνοντας $RMSE = 0.55, MAE = 0.36$ και $R^2 = 0.68$ στα *testing δεδομένα*, αποτέλεσμα που κρίνεται σχετικά ικανοποιητικό, προς το παρών, λαμβάνοντας υπόψη μας την απλότητα που χαρακτηρίζει τον συγκεκριμένο *regressor*. Όπως προκύπτει και ακολούθως από το *Figure 25*, το μοντέλο αυτό είναι αρκετά "μετριοπαθές", μιας και προβλέπει ικανοποιητικά μικρές και μεσαίες τιμές της ζήτησης, ενώ αντιθέτως φαίνεται να αδυνατεί να προβλέψει επιτυχώς έντονα peaks δυστυχώς. Τέλος αξίζει να σημειωθεί ένα από τα βασικά πλεονεκτήματα χρήσης του *RF αλγο-*

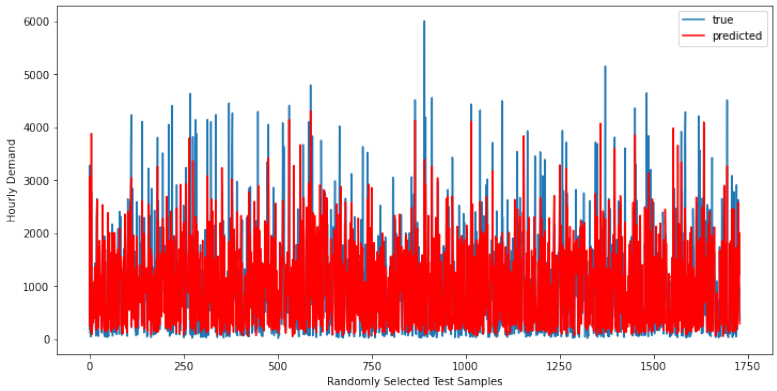


Figure 25: Επίδοση του *Random Forest Regressor* στο *Test Set*

ρίθμου, το οποίο σχετίζεται με την εύρεση της σημαντικότητας των χαρακτηριστικών που συμμετέχουν στην φάση της εκπαίδευσης του. Στην περίπτωση μας, η μέθοδος εντοπίζει ως πιο σημαντικά χαρακτηριστικά τα ακολουθα, σε φθίνουσα σειρά σπουδαιότητας: *Is_ Winter*, *Temprature*, *Humidity*, *Pressure*, *Wind_ Speed*, *Is_ Evening*, *Is_ Afternoon*, *Is_ Weekday*, *Is_ Morning*, *Is_ Night*. Το αντίστοιχο διάγραμμα, σε συνδυασμό με τα αντίστοιχα scores των *top-10* features, εμπεριέχονται στο Python Notebook, μαζί με τον κώδικα όλης της υπόλοιπης εργασίας.

3.1.2 Τεχνικές Βαθιάς Μάθησης (GRU, LSTM, BiLSTM)

Σε αντίθεση με την προηγούμενη μέθοδο, η οποία αποτελεί μία αρκετά δημοφιλή τεχνική *κλασικής μηχανικής μάθησης*, στην συγκεκριμένη ενότητα μελετούμε την επίδοση τριών αρκετά δημοφιλών μοντέλων *βαθιάς μηχανικής μάθησης*, τα οποία τα τελευταία χρόνια απολαμβάνουν ραγδαίας αποδοχής, μιας και χρησιμοποιούνται σε μία πληθώρα εφαρμογών, επιτυγχάνοντας εξαιρετικές επιδόσεις στην πλειονότητα των περιπτώσεων. Ο λόγος για τα

[17]Εν αντιθέσει με τα μοντέλα πρόβλεψης που ακολουθούν, ο αλγόριθμος *Random Forest* γενικότερα λειτουργεί με *vectorized δεδομένα*, χωρίς να έχει δυνατότητα να κάνει track ακολουθιακά patterns. Αυτό ακριβώς είναι και το μειονέκτημα του, το οποίο τον υποχρεώνει σε αυτή την μέτρια επίδοση που παρουσιάζει. Παρόλο αυτά, όπως είναι αναμενόμενο, προκύπτει ότι τα πηγαίνει πολύ πιο καλά σε σχέση με ένα κλασικό *feed-forward νευρωνικό δίκτυο*, για το ίδιο πρόβλημα.

Gated Recurrent Unit (*GRU*), *Long Short Term Memory* (*LSTM*) και *Bidirectional Long Short Term Memory* (*BiLSTM*).

Γενικά μιλώντας, τα προαναφερθέντα τρία μοντέλα ανήκουν στην κατηγορία των *Reccurent Neural Networks* (*RNN*) και ως εκ τούτου είναι σχεδιασμένα να λαμβάνουν υπόψη τους την υφιστάμενη ακολουθιακή δομή των δεδομένων εισόδου. Αυτός ακριβώς είναι και ο λόγος που δικαιολογεί την ευρεία χρήση τους στις μέρες μας, σε προβλήματα *time-series forecasting*. Επιπλέον, δεδομένου του γεγονότος ότι τα τρία αυτά μοντέλα έχουν κάποιες μικρές δομικές διαφορές (βλ. *Figure 26*[18]), οι οποίες δεν είναι της παρούσης να συζητηθούν αναλυτικά, αναμένουμε σε γενικές γραμμές να επιτυγχάνουν παραπλήσιες επιδόσεις. Ακολούθως συζητούμε λεπτομέρειες σχετικά με την δομή των μοντέλων που επιλέξαμε, η οποία είναι κοινή και στις τρεις περιπτώσεις, καθώς και την διαδικασία που ακολουθήσαμε κατά την *φάση εκπαίδευσής* τους.

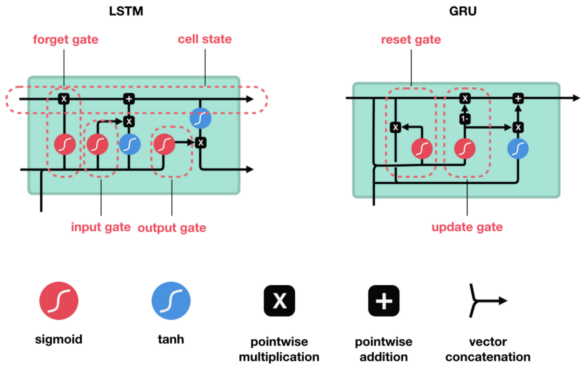


Figure 26: *LSTM νευρώνας* vs *GRU νευρώνας*

Αρχικά και τα τρία μοντέλα αποτελούνται 32 νευρώνες στο *input layer* τους και έναν νευρώνα στο *output layer* τους. Ως *optimizer* έχει επιλεγεί η *μέθοδος Adam*, ενώ θέλοντας να τα είναι να μοντέλα μας πιο robust σε αλλαγές έχουμε εισάγει *Dropout*, με βάση το οποίο σε κάθε *εποχή* κατά την φάση εκπαίδευσης, το 20% τυχαία επιλεγμένων νευρώνων δεν συμμετέχουν στην διαδικασία. Επιπλέον, με στόχο να αποτρέψουμε όσο είναι δυνατό την πιθανότητα να συμβεί *overfitting* έχει χρησιμοποιηθεί *Early-Stopping* με *patience* 10 steps[19], κρατώντας εν τέλει το μοντέλο με το χαμηλότερο *validation loss* εντός του συγκεκριμένου time-window 10 *εποχών*.

Σχετικά με την διαδικασία εκπαίδευσης, δεδομένης της χρήσης *Early-Stopping*, επιλέξαμε να πραγματοποιείται σε 50 *εποχές* το πολύ, με *batch size* = *32* και το 10% των διαθέσιμων δεδομένων χρησιμοποιείται ως *validation set* κάθε φορά, χωρίς να πραγματοποιείται προφανώς κάποιου είδους *shuffling*. Επιπλέον επιλέξαμε τα μοντέλα να πραγματοποιούν προβλέψεις, βασιζόμενα στα τελευταία 30 (χρονικά διατεταγμένα) δείγματα, έπειτα από πειραματισμό, ώστε να επιτυγχάνεται η καλύτερη δυνατή επίδοση αυτών. Εν κατακλείδι, σε αυτό το κομμάτι της εργασίας χρησιμοποιήσαμε το *Keras* της Python, δίνοντας στα

[18] Στο συγκεκριμένο *Figure* παρουσιάζονται μόνο οι υφιστάμενες δομικές διαφορές μεταξύ ενός *GRU νευρώνα* και *LSTM νευρώνα*, μιας και ο *BiLSTM νευρώνας* διαθέτει απλώς δύο επίπεδα (forward και backward διάδοσης), έναντι ενός, με *LSTM νευρώνες*.

[19] Το *patience* έχει τεθεί σχετικά μεγάλο καθώς παρατηρήσαμε ότι σε διαδοχικές εποχές υπήρχαν αισθητές διακυμάνσεις στο *validation loss*. Με αυτό τον τρόπο θελήσαμε να δώσουμε χώρο και χρόνο στα μοντέλα μας για την πιθανή επίτευξη καλύτερων επιδόσεων, πράγμα που κατορθώθηκε και στην πράξη.

μοντέλα μας ως είσο *3D-arrays* διαστάσεων (*#samples, #past_ timesteps, #features*).

Ακολούθως παρουσιάζονται οι επιδόσεις των προαναφερθέντων μοντέλων *βαθιάς μηχανικής μάθησης*, σε μορφή πίνακα, με βάση ορισμένες κλασικές μετρικές για την αξιολόγηση αποτελεσμάτων παλινδρόμησης.

Μοντέλο	RMSE	MAE	$R^2 Score$
GRU	0.1190	0.0797	0.9822
LSTM	0.1358	0.0885	0.9769
BiLSTM	0.1377	0.0882	0.9762

Εύκολα γίνεται κατανοητό ότι και τρία αυτά μοντέλα, όπως είχαμε επισημάνει νωρίτερα, είναι λίγο πολύ αναμενόμενο να επιτυγχάνουν παρόμοιες επιδόσεις. Φυσικά καλύτερο εκ των τριών παρουσιάζεται να είναι το *GRU*, το οποίο διαθέτει το πιο απλό cell εν συγκρίσει με τα άλλα δύο μοντέλα. Ακολούθως (βλ. *Figure 27*) παρουσιάζονται το διάγραμμα του *train - validation error* κατά την φάση της εκπαίδευσης του συγκεκριμένου μοντέλου, καθώς και το τελικό αποτέλεσμα που επιτυγχάνει στο *test set*. Για εξοικονόμηση χώρου, τα αντίστοιχα διαγράμματα των υπόλοιπων μοντέλων παραλείπονται από την αναφορά, όμως είναι ορατά στο επισυναπτόμενο Python Notebook.

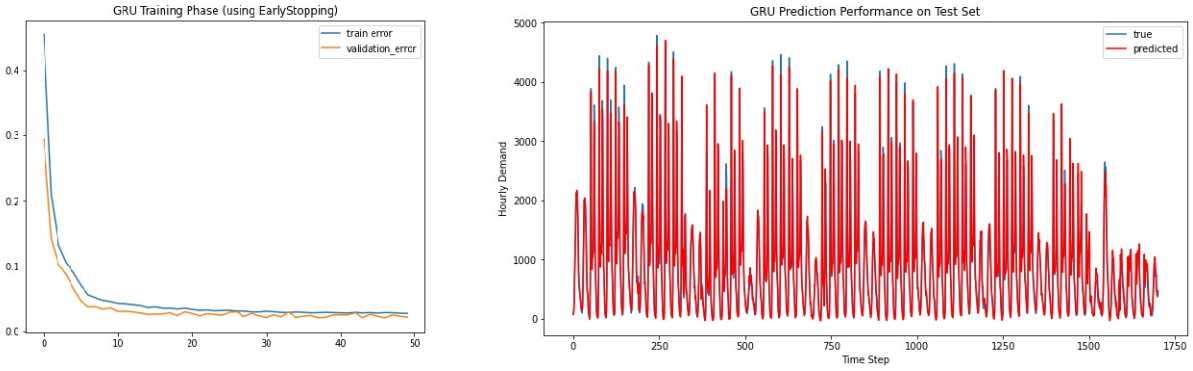


Figure 27: *GRU*: Φάση εκπαίδευσης & επίδοση

3.2 Συσταδοποίηση (Clustering)

Αλλάζοντας τώρα όχθη, στην συγκεκριμένη υποενότητα επικεντρωνόμαστε σε *τεχνικές μηχανικής μάθησης* σε επίπεδο σταθμών. Ειδικότερα ενδιαφερόμαστε για την *γεωγραφική συσταδοποίηση* (*spatial clustering*), ανάλογα με την τοποθεσία τους, προσπαθώντας να δημιουργήσουμε ομάδες από συστάδες κοντινών σταθμών. Προς αυτή την κατεύθυνση, εφαρμόζουμε δύο κλασικές *μεθόδους συσταδοποίησης*, τον *K-Means* και τον *DBSCAN*.

3.2.1 K-Means Μέθοδος

Η μέθοδος *K-Means* αποτελεί ίσως την πιο διαδεδομένη και ταυτόχρονα την πιο απλή *τεχνική συσταδοποίησης*. Στόχος της είναι η δημιουργία ξένων μεταξύ τούς συστάδων, όπου κάθε δείγμα τοποθετείται σε εκείνη από την οποία απέχει την μικρότερη απόσταση από το κέντρο της, έχοντας εκ των προτέρων θεωρήσει προφανώς κάποια μετρική απόστασης. Αυτός είναι και ο λόγος που η συγκεκριμένη μέθοδος προϋποθέτει την ύπαρξη *Ευκλείδειου χώρου* χαρακτηριστικών αποκλειστικά. Ταυτόχρονα, απαραίτητη προϋπόθεση

στην συγκεκριμένη μέθοδο αποτελεί η εκ των προτέρων γνώση του αριθμού των συστάδων. Κάτι τέτοιο φυσικά καθίσταται αδύνατο σε προβλήματα του πραγματικού κόσμου, όπως αυτό που μελετάμε στο πλαίσιο της συγκεκριμένης εργασίας. Υπάρχει φυσικά μια πληθώρα τεχνικών για τον προσδιορισμό του βέλτιστου αριθμού συστάδων, κατά προσέγγιση.

Μία από αυτές είναι η *μέθοδος Elbow*, με την οποία καθίσταται δυνατός ο προσδιορισμός του κατάλληλου αριθμού συστάδων. Ειδικότερα, με βάση την συγκεκριμένη μέθοδο, απεικονίζεται το ποσοστό της συνολικής διασποράς που εξηγείται, συναρτήσει του αριθμού συστάδων, ο οποίος κυμαίνεται εντός κάποιου προκαθορισμένου εύρους. Ως επιθυμητός αριθμός συστάδων ορίζεται εκείνος όπου προσθέτοντας ακόμα μια επιπλέον συστάδα, δεν κερδίζουμε τίποτα το ιδιαίτερο στο ποσοστό "ερμήνευσης" της συνολικής διασποράς. Στην περίπτωση μας, με βάση το *διάγραμμα Elbow* που προκύπτει (βλ.*Figure 28*), θα επιλέγαμε ως $k^* = 3$.

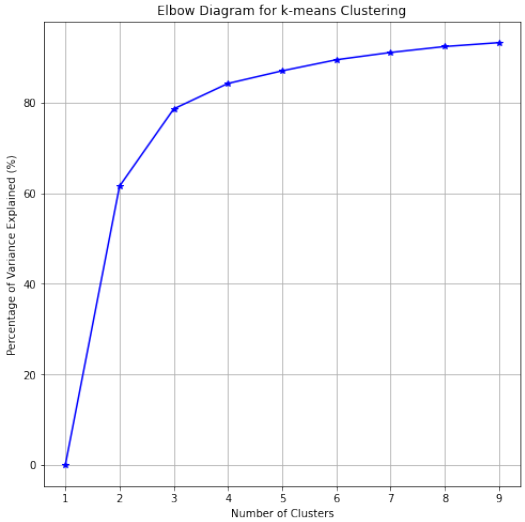


Figure 28: Διάγραμμα *Elbow*

Εν συνεχεία, παρουσιάζονται ακολούθως οι συστάδες σταθμών που δημιουργούνται για $k = 2, 3, 4$ (βλ. *Figure 29*), επαληθεύοντας και οπτικά ότι το αποτέλεσμα που φαίνεται πιο λογικό είναι για $k = 3$. Όπως καθίσταται φανερό, η μέθοδος *K-Means* διαχωρίζει τους

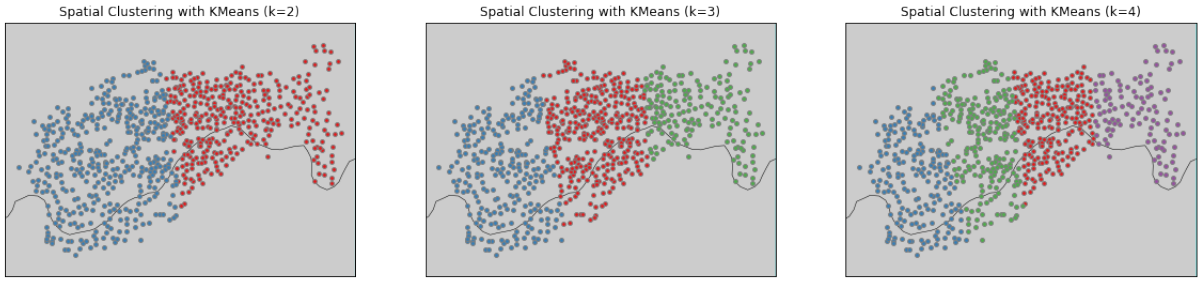


Figure 29: Αποτέλεσμα συσταδοποίησης της μεθόδου *K-Means*

σταθμούς ενοικίασης ποδηλάτων σε περίπου ισοπαχή slices, ανάλογα με το γεωγραφικό τους μήκος. Καθοριστικό ρόλο διαδραματίζει η μορφή του σχήματος που δημιουργούν οι σταθμοί στον επίπεδο, καθώς σε περίπτωση διαφορετικού σχήματος το αποτέλεσμα ενδεχόμενα να ήταν αρκετά διαφορετικό. Αυτό φυσικά είναι γνωστό μειονέκτημα της συγκεκριμένης μεθόδου, με το αποτέλεσμα που παράγεται να μην κρίνεται ιδιαίτερα *informative*.

3.2.2 DBSCAN Μέθοδος

Εάν και στην βιβλιογραφία που μελετήσαμε, σε προβλήματα σχετικά με *BSS*, η μέθοδος *K-Means* για την *γεωγραφική συσταδοποίηση* των σταθμών ήταν αρκετά δημοφιλείς, στην συγκεκριμένη περίπτωση δεν φαίνεται να οδηγεί σε κάποιο πραγματικά ενδιαφέρον αποτέλεσμα. Για τον λόγο αυτό, δοκιμάζουμε την *DBSCAN μέθοδο*, ευελπιστώντας να οδηγηθούμε σε πιο αξιόλογα συμπεράσματα.

Πιο συγκεκριμένα, η μέθοδος *DBSCAN* ανήκει στην οικογένεια των *τεχνικών συσταδοποίησης με βάση την πυκνότητα*. Οι τρεις βασικές παράμετροι που χαρακτηρίζουν την συγκεκριμένη μέθοδο είναι *(α)* η μέγιστη δυνατή απόσταση μεταξύ δύο points, ώστε να θεωρούνται εκείνα εντός της ίδιας γειτονιάς (*eps*), *(β)* ο αριθμός των samples σε μία γειτονιά ώστε κάποιο σημείο να θεωρείται ως *core point* (*min_samples*) και *(γ)* η μετρική απόστασης που έχει επιλεχθεί. Σχετικά με την τελευταία παράμετρο, προκειμένου να προκύψει ένα ορθότερο αποτέλεσμα χρησιμοποιώντας την μέθοδο *DBSCAN*, επιλέξαμε ως μετρική απόστασης την *Haversine distance*, η οποία γενικά αποτελεί την συντομότερη απόσταση μεταξύ δύο σημείων στην επιφάνεια μίας σφαίρας. Με αυτόν τον τρόπο, προσπαθούμε να οδηγηθούμε σε πιο ουσιαστικά αποτελέσματα, απεικονίζοντας τις κλάσεις που διαθέτουν *min_samples* κοντινότερους γείτονες, εντός ακτίνας *eps* μέτρων.

Ακολούθως (βλ. *Figure 30*) παρουσιάζονται, έπειτα από την δοκιμή πολλών και διαφορετικών τιμών για τις παραμέτρους *eps*, *min_samples*, τα καλύτερα οπτικά αποτελέσματα που προέκυψαν, διαμοιράζοντας τους περισσότερους από τους σταθμούς ενοικίασης στην κατάλληλη κλάση. Αναμενόμενα, η μεταβολή των τιμών των δύο αυτών παραμέτρων, οδηγεί σε αρκετά διαφορετικό αποτέλεσμα, καθώς τροποποιείται ο τρόπος με τον οποίο ορίζουμε το μέγεθος και την πυκνότητα της εκάστοτε γειτονιάς. Έπειτα από διαδοχικές δοκιμές που πραγματοποιήσαμε, στην πράξη, παρατηρήσαμε ότι όσο μεγαλώνει η παράμετρος *eps* τόσο μικραίνει το πλήθος των διαφορετικών συστάδων που εντοπίζονται, τείνοντας στον σχηματισμό μίας και μοναδικής καθολικής κλάσης. Αντιθέτως, όσο μεγαλώνει η τιμή της παραμέτρου *min_samples* τόσο η μέθοδος γίνεται πιο επιλεκτική, δυσκολευόμενη να δημιουργήσει πολλές και μεγάλες συστάδες. Φυσικά να σημειωθεί ότι, λαμβάνοντας υπόψιν τον τρόπο λειτουργίας της μεθόδου, αναμένεται εξ' αρχής τέτοιου είδους συμπεριφορά και πειραματικά.

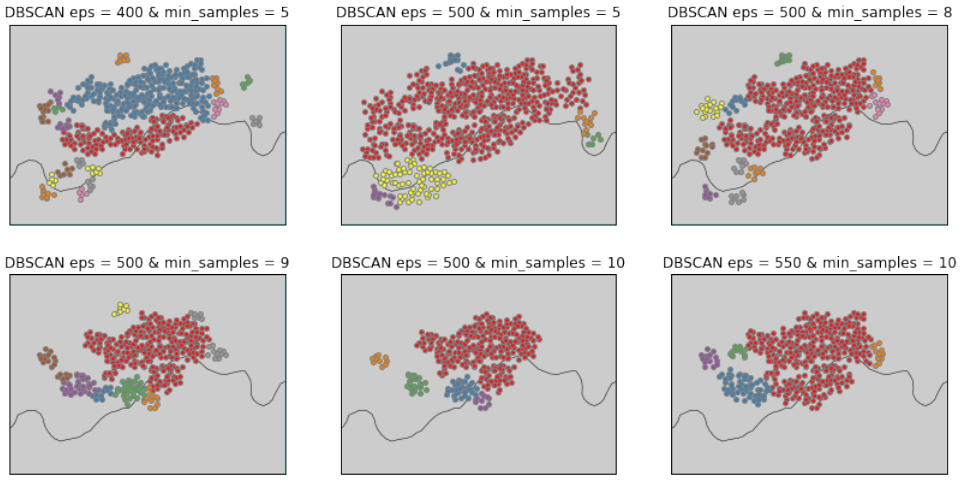


Figure 30: Αποτέλεσμα συσταδοποίησης της μεθόδου *DBSCAN*

,

Τέλος, αξίζει να γίνει μία σύντομη αναφορά, συγκρίνοντας τις δύο αυτές μεθόδους που εξετάστηκαν, παρόλο που προαναφέρθηκε εξ' αρχής ότι η μέθοδος *K-Means* δεν καταλήγει σε επαρκώς ποιοτικά αποτελέσματα. Ειδικότερα, όπως είναι αναμενόμενο άλλωστε, με την μέθοδο *DBSCAN* ορισμένοι σταθμοί χάνονται καθώς εκλαμβάνονται ως σημεία θορύβου από τον αλγόριθμο, εν αντιθέσει με την μέθοδο *K-Means* που ως γνωστόν όλα τα σημεία-σταθμοί κατατάσσονται εν τέλει σε κάποια συστάδα, όπου για την περίπτωση μας έχουν περίπου το ίδιο σχήμα και μέγεθος. Τέλος, σε ορισμένες περιπτώσεις, η μέθοδος *DBSCAN* δύναται να διαχωρίσει σταθμούς που βρίσκονται σε διαφορετική όχθη του *ποταμού Τάμεση*, εν αντιθέσει με την μέθοδο *K-Means*, όπου ανεξαρτήτως περίπτωσης, τέτοιου είδους σταθμοί με παραπλήσιο γεωγραφικό πλάτος κατατάσσονται στην ίδια συστάδα.

4 Σύνοψη Αποτελεσμάτων & Μελλοντικές Κατευθύνσεις

Στο τελευταίο κεφάλαιο της παρούσας εργασίας, καταγράφουμε να σημαντικότερα αποτελέσματα που καταφέραμε να εξάγουμε, σε συνδυασμό με ορισμένες χρήσιμες μελλοντικές κατευθύνσεις, οι οποίες δεν καλύπτονται στην συγκεκριμένη εργασία, ως προέκταση των λεπτομερειών που έχουμε συζητήσει αναλυτικά μέχρι στιγμής.

4.1 Σύνοψη Αποτελεσμάτων & Σχολιασμός

Με βάση όλη την ανάλυση που προηγήθηκε (*Κεφάλαιο 2, 3, 4*), καταλήξαμε σε ορισμένα χρήσιμα συμπεράσματα-αποτελέσματα, τα οποία διατυπώνονται συνοπτικά ακολούθως, συνοψίζοντας την συμβολή μας στην ανάλυση του συγκεκριμένου συνόλου δεδομένων. Ειδικότερα εντοπίσαμε ότι

- Η ωριαία ζήτηση συσχετίζεται ελαφρώς αρνητικά με την υγρασία, ενώ αντιθέτως συσχετίζεται ελαφρώς θετικά με την θερμοκρασία. Το αρχικό αποτέλεσμα που προέκυψε σίγουρα μας ξαφνιάζει, μιας και δεν είναι καθόλου αναμενόμενο. Στον αντίποδα, το δεύτερο αποτέλεσμα είναι απόλυτα λογικό. Επιπλέον η μέση διάρκεια διαδρομής συσχετίζεται ελαφρώς θετικά με την θερμοκρασία, παρομοίως, πράγμα που είναι εξ' αρχής αναμενόμενο.
- Χωρίζοντας καταλλήλως τις διαδρομές σε εποχές, ανάλογα με τον μήνα που έγιναν, οι κατανομές της ζήτησης χειμώνα και άνοιξης είναι παρόμοιες, ενώ κάτι αντίστοιχο συμβαίνει για την περίοδο του καλοκαιριού και του φθινοπώρου. Επίσης, στην δεύτερη περίπτωση παρατηρείται μεγαλύτερο εύρος τιμών, πράγμα που συνδιάζεται με υψηλότερα επίπεδα ζήτησης.
- Κατά τις ημέρες του έτους που είναι καθημερινές, η ζήτηση είναι σημαντικά πιο αυξημένη, λαμβάνοντας τιμές σε μεγαλύτερο εύρος, σε σχέση με την περίπτωση των Σαββατοκύριακων ή επίσημων αργιών.
- Η κατανομή της μηνιαίας μέσης ζήτησης είναι σχεδόν συμμετρική, παρουσιάζοντας peak τους θερινούς μήνες. Αντιθέτως, η μηνιαία κατανομή της διάρκειας διαδρομής είναι σχεδόν ομοιόμορφη, υποδηλώνοντας ότι δεν επηρεάζεται από την εκάστοτε εποχή, η οποία με την σειρά της καθορίζεται από τις αντίστοιχες καιρικές συνθήκες που επικρατούν.

- Τις καθημερινές παρουσιάζονται δύο έντονα peaks στην ζήτηση, ανεξαρτήτου εποχής, μεταξύ 07.00 - 10.00 και 16.00 - 19.00. Προφανώς κάτι τέτοιο δικαιολογείται από το γεγονός ότι τότε η πλειοψηφία του κόσμου μεταβαίνει και επιστρέφει από το χώρο εργασίας τους, αντίστοιχα. Αντιθέτως τα Σαββατοκύριακα ή τις επίσημες αργίες η ζήτηση λαμβάνει πολύ πιο περιορισμένες τιμές, ενώ είναι αρκετά πιο smoothly, χωρίς να εμφανίζονται έντονα peaks μικρού εύρους.

- Το επίπεδο ζήτησης από και προς σταθμούς ενοικίασης που ανήκουν στην ζώνη A, δηλαδή εκείνους με πλησιέστερο σταθμό μετρό στην ζώνη A (κεντρικό Λονδίνο) έχουν σημαντικά υψηλότερη ζήτηση από εκείνη των υπολοίπων σταθμών. Έπειτα ακολουθούν οι σταθμοί της ζώνης B και τέλος απειροελάχιστη είναι η ζήτηση σε σταθμούς ζώνης C. Καθίσταται λοιπόν σαφές, ότι οφείλουμε να εστιάσουμε την προσοχή μας κυρίως σε σταθμούς της ζώνης A και B.

- Οι σταθμούς ενοικίασης που είναι πολύ κοντά (*Very Close*, $d \leq 300m$) και κοντά (*Close*, $300m < d \leq 700m$) στον πιο κοντινό σε αυτούς σταθμό μετρό, παρουσιάζουν παρόμοια επίπεδα ζήτησης από και προς αυτούς, τα οποία είναι αρκετά υψηλότερα σε σχέση με την ζήτηση από και πρός τους σταθμούς που απέχουν αρκετά από τον πλησιέστερο σταθμό μετρό (*Long*, $d > 700m$).

- Η μέση διάρκεια διαδρομής δεν φαίνεται να επηρεάζεται από την ζώνη και την απόσταση από τον πλησιέστερο σταθμό μετρό των σταθμών ενοικίασης αφετηρίας και προορισμού.

- Οι δέκα δημοφιλέστερες διαδρομές, τόσο τις καθημερινές όσο και τα Σαββατοκύριακα, αφορούν σταθμούς στο κεντρικό και βορειοανατολικό Λονδίνο, όπου στην πλειονότητα τους οι σταθμοί αφετηρίας και προορισμού ταυτίζονται. Σε κάθε άλλη περίπτωση, οι σταθμοί αφετηρίας και προορισμού βρίσκονται σε πολύ κοντινή απόσταση και η διαδρομή είναι ιδιαίτερα σύντομη.

- Η μέθοδος *Google's PageRank* δημιουργεί ένα διαφορετικό ranking μεταξύ των σταθμών, σε σχέση με εκείνο που προκύπτει λαμβάνοντας υπόψη απλώς την δημοφιλία των σταθμών[20], μιας και επιβραβεύει σταθμούς που συνδέονται με άλλους "σημαντικούς" σταθμούς, καταλήγοντας σε ένα πιο διαισθητικά αποδεκτό αποτέλεσμα.

- Έπειτα από την δοκιμή μιας πληθώρας κλασικών μεθόδων μηχανικής μάθησης, ο *Random Forest (RF) Regressor* παρουσίαζε την καλύτερη επίδοση στο πρόβλημα της *πρόβλεψης της ωριαίας ζήτησης*, η οποία όμως δεν είναι αρκετά ικανοποιητική, εξαιτίας την αδυναμίας του να κάνει track ακολουθιακά δεδομένα. Επιπλέον, η συγκεκριμένη μέθοδος παρουσιάζει μια ιδιαίτερα "μετριοπαθή" συμπεριφορά, προβλέποντας ικανοποιητικά χαμηλά και μέσα demands, παρουσιάζοντας όμως αδυναμία στην πρόβλεψη υψηλότερων demands.

- Οι *τεχνικές βαθιάς μηχανικής μάθησης* που δοκιμάστηκαν, ονομαστικά *GRU, LSTM, BiLSTM*, παρουσιάζουν πολύ καλύτερες επιδόσεις, οι οποίες είναι σχετικά παρόμοιες και για τα τρία μοντέλα, ξεχωρίζοντας στο συγκεκριμένο πρόβλημα η απλούστερη τεχνική, το *GRU* μοντέλο.

- Για το πρόβλημα της συσταδοποίησης των σταθμών, ανάλογα με την τοποθεσία τους, δυστυχώς η μέθοδος *K-Means* δεν παρουσιάζει αρκετά ικανοποιητικά αποτελέσματα,

[20] Όπως είχε αναφερθεί και προηγουμένως, η δημοφιλία του εκάστοτε σταθμού κρίνεται με βάση το συνολικό πλήθος διαδρομών που έχουν ξεκινήσει ή καταλήξει στο συγκεκριμένο σταθμό.

μιας και χωρίζει τους σταθμούς σε ομάδες με κοντινό γεωγραφικό μήκος απλώς. Κάτι τέτοιο, φυσικά, οφείλεται στο σχήμα που δημιουργεί το σύνολο των σταθμών, καθώς η συγκεκριμένη μέθοδος είναι ιδιαίτερη ευαίσθητη σε αυτό, σε σχέση με το προκύπτον αποτέλεσμα στο οποίο θα καταλήξει. Αντιθέτως με την περίπτωση της τεχνικής *K-Means*, η μέθοδος *DBSCAN* οδηγεί σε πιο χρήσιμα αποτελέσματα, μιας και πρόκειται για μέθοδο *συσταδοποίησης με βάση την πυκνότητα*. Ως αποτέλεσμα, δημιουργεί ομάδες σταθμών οι οποίοι εντός μιας προκαθορισμένης ακτίνας, διαθέτουν πάνω από ένα πλήθος σταθμών.

4.2 Μελλοντικές Κατευθύνσεις

Έπειτα από την επιτέλεση διεξοδικής μελέτης και την ενασχόλησής μας, στο πλαίσιο της συγκεκριμένης εργασίας, σε μία προσπάθεια να εξαχθεί όσο το δυνατό περισσότερη γνώση, εντοπίσαμε ορισμένα σημεία που μπορούν να αποτελέσουν βασικές μελλοντικές κατευθύνσεις, επεκτείνοντας το περιεχόμενο της παρούσας εργασίας.

Προκείμενου, λοιπόν, να είμαστε σε θέση να κατανοήσουμε καλύτερα την συνολική εικόνα και να εξάγουμε περισσότερο χρήσιμα και ενδιαφέροντα insights, απαιτείται η χρησιμοποίηση επιπλέον συνόλων δεδομένων, όπως για παράδειγμα δεδομένων που σχετίζονται με την ταυτότητα του εκάστοτε ενοικιαστή-μέλους. Προφανώς το ενδιαφέρον των bike vendors αναμένεται να είναι μεγάλο, για τέτοιου είδους αποτελέσματα, καθώς με αυτό τον τρόπο θα καθίσταται δυνατή η αποτελεσματικότερη διαχείριση των *BSS*, πράγμα που συνεπάγεται άμεσα την αύξηση αύξηση των κερδών τους. Επιπρόσθετα, όπως είχε αναφερθεί και στο εισαγωγικό κεφάλαιο, βασικός σκοπός προβλημάτων που σχετίζονται με *BSS*, αποτελεί ο προσδιορισμός κάποιας όσο το δυνατό καλύτερης εφικτής λύσης για το *πρόβλημα του rebalancing* ποδηλάτων σε σταθμούς αρκετά δημοφιλείς, μιας και εκ φύσεως ένα τέτοιο πρόβλημα είναι ιδιαίτερα δύσκολο. Ως μια εκ μία εκ των πολλά υποσχόμενων εναλλακτικών, προβάλει η χρήση μεθόδων *βαθιάς μηχανικής μάθησης*, συνδυάζοντας δεδομένα από πολλές και διαφορετικές πηγές, με την αξιοποίηση ενδεχομένως και κατανεμημένων συστημάτων, προκείμενου να καθίσταται δυνατή η διαχείριση και η επεξεργασία δεδομένων μεγάλης κλίμακας. Μεγάλη πρόκληση αποτελεί, επίσης, η δημιουργία ενός αποδοτικού συστήματος *rebalancing* που να λειτουργεί σε πραγματικό χρόνο και να δύναται να τροποποιεί τις αποφάσεις που λαμβάνει, δυναμικά, ανάλογα με τις τρέχουσες συνθήκες που επικρατούν, επιτυγχάνοντας όσο το δυνατό καλύτερη εμπειρία για τα μέλη, καθώς και περισσότερα κέρδη για τις εταιρίες ενοικίασης.

Επίσης, μία αρκετά ενδιαφέρουσα περίπτωση που δεν συζητήθηκε στην συγκεκριμένη εργασία θα ήταν η προέκταση της ιδέας περί *γεωγραφικής συσταδοποίησης* σταθμών, σε *συσταδοποίηση με βάση της ροές κίνησης*, χρησιμοποιώντας πιθανώς κάποιον *ιεραρχικό αλγόριθμο*. Ως αποτέλεσμα, θα ήταν δυνατή η δημιουργία ομάδων, οι οποίες θα αποτελούνταν από σταθμούς με παρόμοιες ροές, προσφέροντας σημαντική πληροφορία στις εταιρίες ενοικίασης ποδηλάτων, σχετικά με κλάσεις σταθμών που αντιστοιχούν σε υψηλότερες ροές κίνησης και απαιτούν συνεπώς μεγαλύτερη προσοχή. Η συγκεκριμένη κατεύθυνση συνδέεται άμεσα με το πρόβλημα του *rebalancing*, που αποτελεί ίσως το βασικότερο πρόβλημα προς επίλυση σε *BSS*, όπως έχουμε συζητήσει επανειλημμένα. Στην υπάρχουσα βιβλιογραφία, έχουμε ήδη εντοπίσει ένα μικρό πλήθος εργασιών προς αυτή την κατεύθυνση, όπως για παράδειγμα [6].

Τέλος, μία διαφορετική και αρκετά ενδιαφέρουσα προσέγγιση της ωριαίας πρόβλεψης ζήτησης,

θα ήταν με την χρήση κλασικών στατιστικών μεθόδων για *time-series forecasting*, όπως για παράδειγμα το μοντέλο *ARIMA* ή *SARIMA*. Βασική μελλοντική επιδίωξη, θα μπορούσε να αποτελέσει η πειραματική σύγκριση των δύο αυτών οικογενειών μοντέλων, οι οποίες διακατέχονται από εντελώς διαφορετική νοοτροπία, για το συγκεκριμένο πρόβλημα. Παρόλο αυτά, εκ των προτέρων γνωρίζουμε ότι σε περιπτώσεις όπου μεγάλα σύνολα δεδομένων είναι διαθέσιμα, όπως συμβαίνει στην δική μας περίπτωση, οι *τεχνικές βαθιάς μηχανικής μάθησης* επιτυγχάνουν καλύτερα αποτελέσματα. Τέτοιου είδους στατιστικές μέθοδοι, στην πραγματικότητα, είναι χρήσιμες σε περιπτώσεις όπου το υπάρχον dataset είναι δραματικά μικρό, όπως θα συνέβαινε για παράδειγμα εάν μας ενδιέφερε η ημερήσια έναντι της ωριαίας πρόβλεψης ζήτησης.

Παράρτημα Ι: Τεχνικές Λεπτομέρειες περί της Εργασίας

Ο κώδικας της συγκεκριμένης εργασίας βρίσκεται πλήρως στο επισυναπτόμενο Python Notebook, μαζί με αναλυτικά σχόλια, ώστε να καθίσταται δυνατή η κατανόηση των βημάτων που ακολουθήσαμε. Λόγω του μεγάλου όγκου των datasets, σε περίπτωση που είναι επιθυμητό, είναι διαθέσιμα στο ακόλουθο link:

https://drive.google.com/file/d/1paxKVXlOojgOXAfiKJh_tBn1eRYNmhhq/view?usp=sharing

Στο πλαίσιο της συγκεκριμένης εργασίας, επισυνάπτονται συνολικά τρία αρχεία, 1 Notebook με τον απαραίτητο κώδικα σε Python, το Report (παρόν αρχείο) στο οποίο επεξηγείται αναλυτικά η διαδικασία, καθώς και η γενικότερη φιλοσοφία που επιλέξαμε να ακολουθήσουμε, βήμα προς βήμα και τέλος η σύντομη επισκόπηση της εργασίας υπό την μορφή παρουσίασης.

Αναφορές

[1] London Bicycle Hires dataset. [Online:] https://console.cloud.google.com/marketplace/product/greater-london-authority/london-bicycles?filter=solution-type:dataset&filter=category:encyclopedic&id=95374cac-2834-4fa2-a71f-fc033ccb5ce4.

[2] Weather Underground website. [Online:] https://www.wunderground.com/history/daily/gb/london/EGLC/date/2015-1-1.

[3] A. Chakure. Random forest regression. [Online:] https://medium.com/@aaaanchakure/random-forest-and-its-implementation-71824ced454f.

[4] L. Chen, D. Zhang, L. Wang, D. Yang, X. Ma, S. Li, Z. Wu, G. Pan, TMT Nguyen, and J. Jakubowicz. Dynamic Cluster-Based Over-Demand Prediction in Bike Sharing Systems. *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 2016.

[5] Y. Li, Y. Zheng, H. Zhang, and L. Chen. Traffic Prediction in a Bike-Sharing System. New York, NY, USA, 2015. Association for Computing Machinery.

[6] L. Liu, D. Gong, B. Guan, and J. Xiao. Cf-cluster: Clustering bike station based on common flows. In *2017 IEEE 15th Intl Conf on Dependable, Autonomic and Secure Computing, 15th Intl Conf on Pervasive Intelligence and Computing, 3rd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress(DASC/PiCom/DataCom/CyberSciTech)*, pages 356–361, 2017.

[7] P. Mrazovic, J. L. Larriba-Pey, and M. Matskin. A Deep Learning Approach for Estimating Inventory Rebalancing Demand in Bicycle Sharing Systems. In *2018 IEEE 42nd Annual Computer Software and Applications Conference (COMPSAC)*, volume 02, 2018.

[8] E. O'Mahony and D. Shmoys. Data Analysis and Optimization for (Citi)Bike Sharing. In *AAAI*, 2015.

[9] A. Rajaraman, J. Leskovec, and J. Ullman. Mining of Massive Datasets, pages 175 - 212. [Online:] http://infolab.stanford.edu/~ullman/mmds/book0n.pdf.

[10] D. Singhvi, S. Singhvi, P. Frazier, S. Henderson, E. O'Mahony, D. Shmoys, and D. B. Woodard. Predicting Bike Usage for New York City's Bike Sharing System. In *AAAI Workshop: Computational Sustainability*, 2015.

[11] A. Singla, M. Santoni, Gábor Bartók, Pratik Mukerji, Moritz Meenen, and Andreas Krause. Incentivizing Users for Balancing Bike Sharing Systems. In *AAAI*, 2015.

[12] P. Vogel, T. Greiser, and D. C. Mattfeld. Understanding Bike-Sharing Systems using Data Mining: Exploring Activity Patterns. *Procedia - Social and Behavioral Sciences*, 20:514 – 523, 2011.

[13] J. Zhang, X. Pan, M. Li, and P. S. Yu. Bicycle-Sharing System Analysis and Trip Prediction. In *2016 17th IEEE International Conference on Mobile Data Management (MDM)*, volume 1, pages 174–179, 2016.

[14] Διαφάνειες Διαλέξεων και Υλικό Εργαστηρίου Μαθήματος. [Online:] https://eclass.ails.ece.ntua.gr/modules/document/?course=103.