Θέματα Επιστήμης Δεδομένων – Electricity Load Diagrams Dataset

Τεχνική Αναφορά της
εργασίας του μαθήματος
«Θέματα Επιστήμης Δεδομένων».

Ακαδημαϊκό Έτος: 2022 – 2023

Ομάδα:

Μπουμπλίνη Αναστασία (Π19117)
aboublini@gmail.com
ANASTASIA BOUBLINI (p19117@unipi.gr)

Μπριστογιάννης Ιωακείμ (Π19048)
ioakeim13@hotmail.gr
IOAKEIM EL-KHATTAB-BRISTOGIANNIS (p19048@unipi.gr)

Παλιούρα Παρασκευή (Π19129)
paliourp@gmail.com
PARASKEVI PALIOURA (p19129@unipi.gr)

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ
UNIVERSITY OF PIRAEUS

Πίνακας Περιεχομένων

Πρόλογος

Η παρούσα εργασία ανάγεται στην μελέτη και την ανάλυση του συνόλου δεδομένων "Electricity Load Diagrams" από το UCI, το οποίο αποτελείται από 370 χρονοσειρές κατανάλωσης ηλεκτρικού ρεύματος που συλλέχθηκαν από πελάτες ενός παρόχου ενέργειας στην Πορτογαλία κατά το χρονικό διάστημα 2011-2014.

Η συγκεκριμένη εργασία ανήκει στο πεδίο της Μηχανικής Μάθησης (Machine Learning) και πιο συγκεκριμένα ασχολείται με την ανάλυση των δοθέντων δεδομένων. Στόχος της είναι η πρόβλεψη της κατανάλωσης ενέργειας τόσο βραχυπρόθεσμα (εντός τριών ωρών) όσο και μακροπρόθεσμα (εντός τριών ηχερών) και χωρίζεται στα παρακάτω βασικά μέρη:

Ά Μέρος: Προεπεξεργασία Δεδομένων

΄Β Μέρος: Συσταδοποίηση

΄Γ Μέρος: Πρόβλεψη

Η εργασία αναπτύχθηκε σε jupyter notebook, με τη χρήση Python και το περιβάλλον προγραμματισμού που χρησιμοποιήθηκε είναι το Visual Studio Code.

1. Μέρος Ά – Προεπεξεργασία Δεδομένων

Η εξοικείωση με τα δεδομένα αποτελεί απαραίτητο βήμα σε οποιοδήποτε έργο που σχετίζεται με την Επιστήμη Δεδομένων και τη Μηχανική Μάθηση. Είναι το πρώτο και το βασικότερο βήμα που πρέπει να γίνει προτού ο ερευνητής προβεί σε οποιαδήποτε άλλη ενέργεια, καθώς περιλαμβάνει τις προπαρασκευαστικές εργασίες που πρέπει να γίνουν στο σύνολο δεδομένων, ώστε να «καθαριστεί» από περιττές ή εσφαλμένες πληροφορίες, να κανονικοποιηθούν τα δεδομένα και να γίνει η οπτικοποίηση τους.

1.1. Κατανόηση & προετοιμασία των δεδομένων

Πρωτού γίνει οποιαδήποτε ενέργεια σχετικά με την ανάλυση του δοθέντος συνόλου δεδομένων, είναι απαραίτητο να προηγηθεί η απόλυτη κατανόηση του. Για αυτόν το λόγο έχουν εφαρμοστεί κάποιες μέθοδοι που αποσκοπούν στην ανάλυση του συνόλου, όπως επίσης την εξαγωγή κάποιων στατιστικών.

Αρχικά, όπως φαίνεται και στις παρακάτω εικόνες (Εικόνες 1 & 2), όσον αφορά την διάσταση του συνόλου, περιλαμβάνονται 371 στήλες και 140.256 γραμμές. Σχετικά με τους τύπους δεδομένων, οι 370 στήλες είναι τύπου float (οι πελάτες) και μια στήλη είναι τύπου object (χρονοσφραγίδα κατανάλωσης). Επιπλέον, με την περιγραφική στατιστική ανάλυση της Εικόνας 3, μπορούμε να εξάγουμε χρήσιμες πληροφορίες όπως η συνολική κατανάλωση ή η μέγιστη / ελάχιστη τιμή κατανάλωσης κ.α. ανά πελάτη. Τέλος είναι αναγκαίο να προσδιοριστούν οι null τιμές που υπάρχουν και στη συνέχεια, αν υπάρχουν, να αφαιρεθούν από το σύνολο, ώστε να μην επηρεαστούν τα αποτελέσματα των αλγορίθμων μηχανικής μάθησης. Στην προκειμένη περίπτωση, δεν υπάρχουν null τιμές σε καμία στήλη του συνόλου (Εικόνα 4), επομένως δεν χρειάζεται να εφαρμόσουμε κάποια από τις ενέργειες που περιγράφηκαν παραπάνω.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 140256 entries, 0 to 140255
Columns: 371 entries, Datetime to MT_370
dtypes: float64(370), object(1)
memory usage: 397.0+ MB
```

Εικόνα 1: Συνοπτική περιγραφή του συνόλου δεδομένων

	Datetime	MT_001	MT_002	MT_003	MT_004	MT_005	MT_006	MT_007	MT_008	MT_009	...	MT_361	MT_362	MT_363	MT_364	MT_365	MT_366
140251	2014-12-31 23:00:00	2.538071	22.048364	1.737619	150.406504	85.365854	303.571429	11.305822	282.828283	68.181818	...	276.945039	28200.0	1616.033755	1363.636364	29.986962	5.851375
140252	2014-12-31 23:15:00	2.538071	21.337127	1.737619	166.666667	81.707317	324.404762	11.305822	252.525253	64.685315	...	279.800143	28300.0	1569.620253	1340.909091	29.986962	9.947338
140253	2014-12-31 23:30:00	2.538071	20.625889	1.737619	162.601626	82.926829	318.452381	10.175240	242.424242	61.188811	...	284.796574	27800.0	1556.962025	1318.181818	27.379400	9.362200
140254	2014-12-31 23:45:00	1.269036	21.337127	1.737619	166.666667	85.365854	285.714286	10.175240	225.589226	64.685315	...	246.252677	28000.0	1443.037975	909.090909	26.075619	4.095963
140255	2015-01-01 00:00:00	2.538071	19.914651	1.737619	178.861789	84.146341	279.761905	10.175240	249.158249	62.937063	...	188.436831	27800.0	1409.282700	954.545455	27.379400	4.095963

5 rows × 371 columns

Εικόνα 2: Οι τελευταίες πέντε γραμμές του συνόλου δεδομένων

	MT_001	MT_002	MT_003	MT_004	MT_005	MT_006	MT_007	MT_008	MT_009	MT_010	...	MT_361	MT_362
count	140256.000000	140256.000000	140256.000000	140256.000000	140256.000000	140256.000000	140256.000000	140256.000000	140256.000000	140256.000000	...	140256.000000	140256.000000
mean	3.970785	20.768480	2.918308	82.184490	37.240309	141.227385	4.521338	191.401476	39.975354	42.205152	...	218.213701	37607.987537
std	5.983965	13.272415	11.014456	58.248392	26.461327	98.439984	6.485684	121.981187	29.814595	33.401251	...	204.833532	38691.954832
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000
25%	0.000000	2.844950	0.000000	36.585366	15.853659	71.428571	0.565291	111.111111	13.986014	9.677419	...	5.710207	0.000000
50%	1.269036	24.893314	1.737619	87.398374	39.024390	157.738095	2.826456	222.222222	40.209790	40.860215	...	131.334761	24100.000000
75%	2.538071	29.871977	1.737619	115.853659	54.878049	205.357143	4.522329	279.461279	57.692308	61.290323	...	403.283369	54800.000000
max	48.223350	115.220484	151.172893	321.138211	150.000000	535.714286	44.657999	552.188552	157.342657	198.924731	...	852.962170	192800.000000

8 rows × 370 columns

Εικόνα 3: Περιγραφικά στατιστικά του συνόλου δεδομένων

```
Datetime    0
MT_001      0
MT_002      0
MT_003      0
MT_004      0
           ..
MT_366      0
MT_367      0
MT_368      0
MT_369      0
MT_370      0
Length: 371, dtype: int64
```

Εικόνα 4: Έλεγχος για null τιμές στο σύνολο δεδομένων

Όσον αφορά την προετοιμασία των δεδομένων, κρίθηκε απαραίτητη η εξαγωγή πληροφοριών από την στήλη Datetime. Συγκεκριμένα από την συγκεκριμένη στήλη δημιουργήθηκαν οι στήλες “Full Date”, “Time”, “Year”, “Month”, “Day Name” και “Month Name” και στη συνέχεια η στήλη Datetime αφαιρέθηκε από το σύνολο δεδομένων, καθώς περιείχε μεγάλο όγκο πληροφορίας, ο οποίος δεν μπορούσε να χρησιμοποιηθεί εύκολα και ήταν προτιμότερο να διασπαστεί σε επι μέρους κατηγορίες (Εικόνα 5). Επιπλέον οι εγγραφές που αντιστοιχούν στις χρονιές 2011 και 2015 αφαιρέθηκαν από το σύνολο δεδομένων, καθώς οι εγγραφές του 2011 είναι μηδενικές, ενώ στο 2015 παρατηρείται μια μόνο εγγραφή, γεγονότα τα οποία θα αλλοίωναν τα αποτελέσματα των αλγορίθμων μηχανικής μάθησης, επομένως κρίθηκε απαραίτητη η διαγραφή τους.

Datetime
2014-12-31 23:00:00
2014-12-31 23:15:00
2014-12-31 23:30:00
2014-12-31 23:45:00
2015-01-01 00:00:00

Full Date	Time	Year	Day	Month	Day Name	Month Name
2012-01-01	00:00:00	2012	1	1	Sunday	January
2012-01-01	00:15:00	2012	1	1	Sunday	January
2012-01-01	00:30:00	2012	1	1	Sunday	January
2012-01-01	00:45:00	2012	1	1	Sunday	January
2012-01-01	01:00:00	2012	1	1	Sunday	January

Εικόνα 4: Διάσπαση της στήλης Datetime σε επι μέρους στήλες (εξαγωγή πληροφορίας)

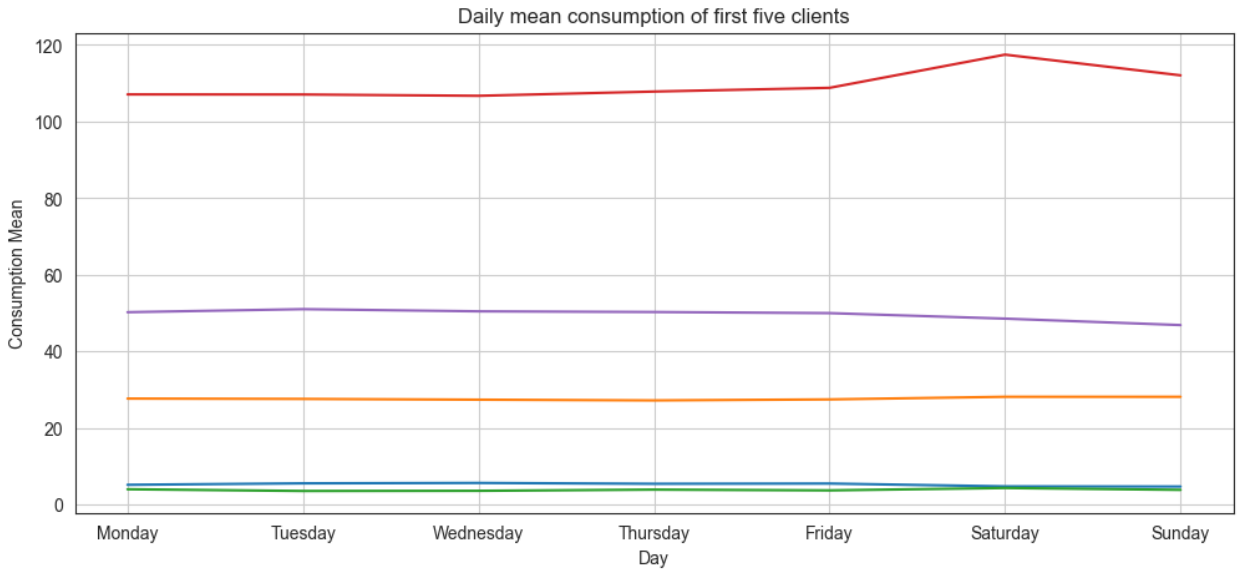
1.2. Οπτικοποίηση των δεδομένων

Οπτικοποίηση των δεδομένων είναι η προβολή της πληροφορίας είτε σε μορφή γραφήματος, οποιουδήποτε είδους, είτε σε μορφή πίνακα. Η τεχνική της οπτικοποίησης στην Αναλυτική Δεδομένων θεωρείται απαραίτητο στάδιο και η σημασία της έγκειται στο ότι οι ερευνητές μπορούν να αφομοιώσουν πιο γρήγορα μεγάλο όγκο οπτικής πληροφορίας και κατ' επέκταση να εξοικειωθούν σε σύντομο χρονικό διάστημα με τα δεδομένα τους. Στην εν λόγω εργασία η οπτικοποίηση περιλαμβάνει:

- Ημερήσια μέση κατανάλωση των πελατών

Day Name	MT_001	MT_002	MT_003	MT_004	MT_005	MT_006	MT_007	MT_008	MT_009	MT_010	...	MT_361	MT_362	MT_363	MT_364	MT_365	MT_366	MT_367	MT_368	MT_369	MT_370
Monday	5.206060	27.702212	4.048278	107.068861	50.245246	187.681865	5.493436	252.766965	55.759511	62.865628	...	286.292936	49352.521231	2461.080765	3881.221651	84.687546	14.896665	563.069057	139.368760	832.961073	11153.529724
Tuesday	5.589246	27.615950	3.592660	107.042834	51.037216	192.333348	5.975913	258.390031	56.375667	59.403822	...	285.044586	49792.860934	2461.933213	3887.895677	88.785304	10.932583	593.634138	159.484447	870.032836	12148.950623
Wednesday	5.692389	27.419029	3.637046	106.703271	50.468645	189.392898	6.063227	257.396372	56.435984	59.788071	...	285.003906	49902.103238	2464.842175	3915.330655	87.586538	15.274682	587.329021	157.737352	860.881892	12070.831899
Thursday	5.474580	27.227917	3.947585	107.813287	50.292338	192.089089	6.367641	256.946243	56.195040	58.967753	...	287.433436	50223.344017	2476.153407	3928.755706	87.731156	11.210603	592.394304	153.888737	872.826043	12321.228199
Friday	5.536439	27.489930	3.750455	108.782453	50.000814	190.210885	6.346993	263.350204	55.449558	60.130474	...	299.003268	50175.173611	2545.178897	3976.859946	88.114647	15.303323	608.067963	153.326123	879.521555	12425.824382
Saturday	4.798202	28.171295	4.372301	117.445772	48.562822	191.508994	6.251495	261.670525	47.842632	48.789174	...	303.059103	50688.154380	2620.103276	3958.823512	87.277149	10.737758	549.098751	68.909922	798.543652	11368.573140
Sunday	4.754673	28.166225	3.885433	112.046189	46.883819	174.643457	5.695781	235.574741	44.953222	43.869641	...	290.454166	50796.204883	2583.180481	3885.768011	86.209056	8.142814	465.585849	50.968374	719.861381	9908.528074

Εικόνα 5: Ημερήσια μέση κατανάλωση κάθε πελάτη

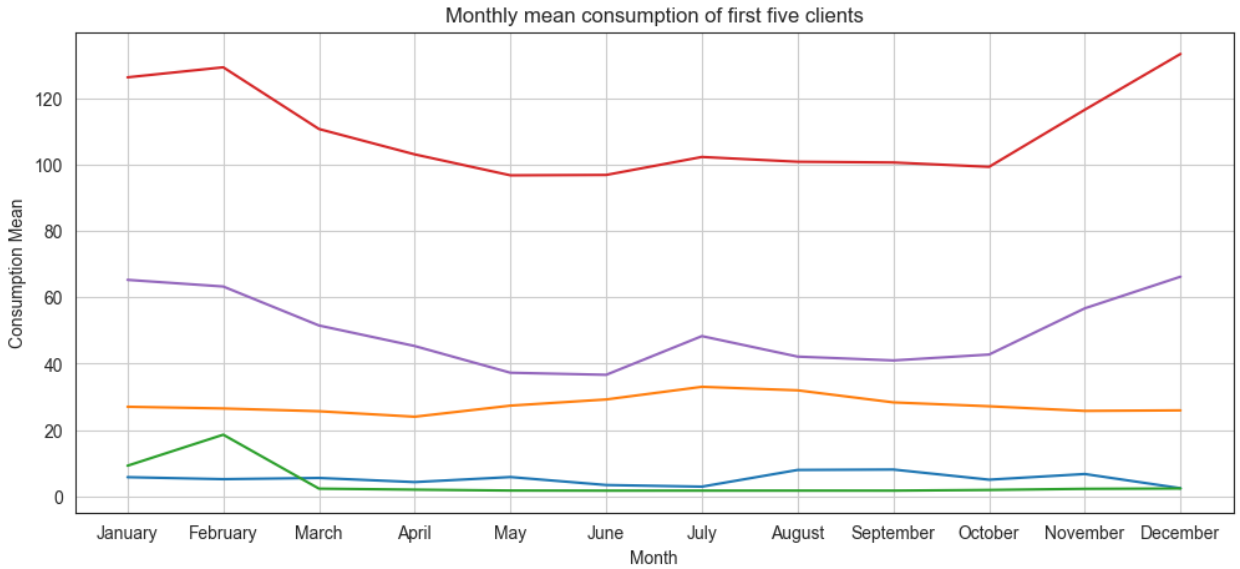


Εικόνα 6: Ημερήσια μέση κατανάλωση κάθε πελάτη (στο γράφημα φαίνονται οι πρώτοι πέντε πελάτες για λόγους ευκρίνειας)

- Μηνιαία μέση κατανάλωση των πελατών

Month Name	MT_001	MT_002	MT_003	MT_004	MT_005	MT_006	MT_007	MT_008	MT_009	MT_010	...	MT_361	MT_362	MT_363	MT_364	MT_365	MT_366
January	5.807884	27.017308	9.259811	126.257121	65.251005	231.644159	4.463951	270.107738	69.533149	67.994494	...	256.913025	34284.162186	2141.659767	2939.559405	69.459778	12.452326
February	5.211599	26.506499	18.606370	129.279003	63.243364	228.734827	4.346507	261.634730	63.433901	61.314174	...	242.480336	31724.301471	2046.748056	2712.530637	75.147794	12.006864
March	5.574631	25.669876	2.354097	110.688352	51.501442	192.241303	4.012946	244.943461	50.146902	54.840034	...	246.100296	33441.487455	2113.644742	3213.149642	74.063900	11.897141
April	4.327793	24.019579	2.048541	103.046899	45.313206	172.291116	3.512264	236.977023	45.247952	53.672466	...	251.162282	37219.814815	2094.693995	3495.941183	74.158729	10.561393
May	5.846547	27.360977	1.778588	96.736175	37.266424	163.207779	2.999817	234.778339	40.608473	50.746353	...	272.123934	48918.615591	2307.034334	4102.372006	82.356082	10.381013
June	3.449456	29.220092	1.742044	96.851005	36.646483	162.620908	2.927541	245.373878	42.086530	42.408652	...	316.788686	60355.706019	2664.664401	4786.887100	94.931250	14.270109
July	2.953976	33.029289	1.750465	102.265504	48.303365	170.604919	17.361174	261.517882	49.606448	54.366112	...	351.886709	77587.992832	3122.422890	5123.813742	111.987308	14.779310
August	7.984775	31.971673	1.757277	100.829673	42.128082	161.804022	15.208408	262.430156	45.378569	45.108418	...	371.344377	82407.515681	3213.549465	4988.058508	101.815976	13.759186
September	8.108168	28.334255	1.750591	100.603169	40.977868	169.778232	5.102080	256.562149	55.033913	50.877763	...	337.628298	70598.576389	3062.901430	4777.643624	111.882364	13.159499
October	5.054821	27.170740	1.956574	99.291395	42.749448	173.855127	3.079026	251.074062	58.737577	62.673310	...	307.863175	51605.365143	2805.987516	4222.010936	104.290502	13.124566
November	6.762167	25.768038	2.300837	116.489997	56.675136	200.766438	4.009707	258.080029	57.787814	67.843613	...	266.631474	37523.171296	2395.695617	3491.832386	72.832464	10.786915
December	2.476524	25.938343	2.393800	133.305605	66.202247	233.489277	4.879752	278.015369	62.304967	63.379557	...	265.794930	34503.897849	2187.390828	3100.325330	72.567648	11.051680

Εικόνα 7: Μηνιαία μέση κατανάλωση κάθε πελάτη

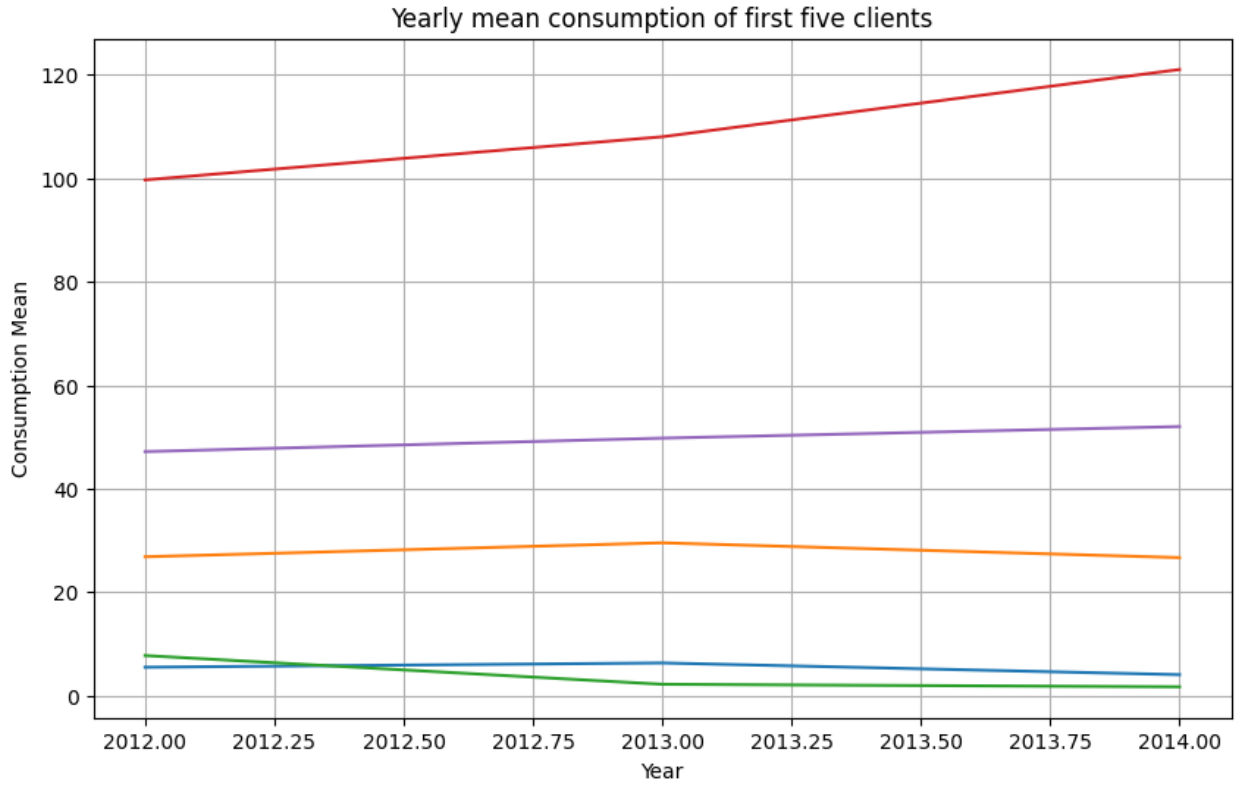


Εικόνα 8: Μηνιαία μέση κατανάλωση κάθε πελάτη (στο γράφημα φαίνονται οι πρώτοι πέντε πελάτες για λόγους ευκρίνειας)

- Ετήσια μέση κατανάλωση των πελατών

Year	MT_001	MT_002	MT_003	MT_004	MT_005	MT_006	MT_007	MT_008	MT_009	MT_010	...	MT_361	MT_362	MT_363	MT_364	MT_365	MT_366
2012	5.496698	26.842841	7.750324	99.668929	47.163530	185.883577	4.806117	254.049872	55.001017	53.203680	...	295.679096	53383.956626	2798.317376	4071.638516	101.108481	12.780865
2013	6.324027	29.535317	2.200142	108.006935	49.765773	192.505504	7.414558	260.241782	54.035348	60.812681	...	291.472436	48697.308790	2461.028505	4028.794365	77.287979	13.106546
2014	4.058161	26.678553	1.709477	121.009091	52.002346	186.390129	5.863605	251.135499	50.822087	54.772119	...	285.485661	48306.669521	2287.827775	3656.566521	83.157561	11.181687

Εικόνα 9: Ετήσια μέση κατανάλωση κάθε πελάτη

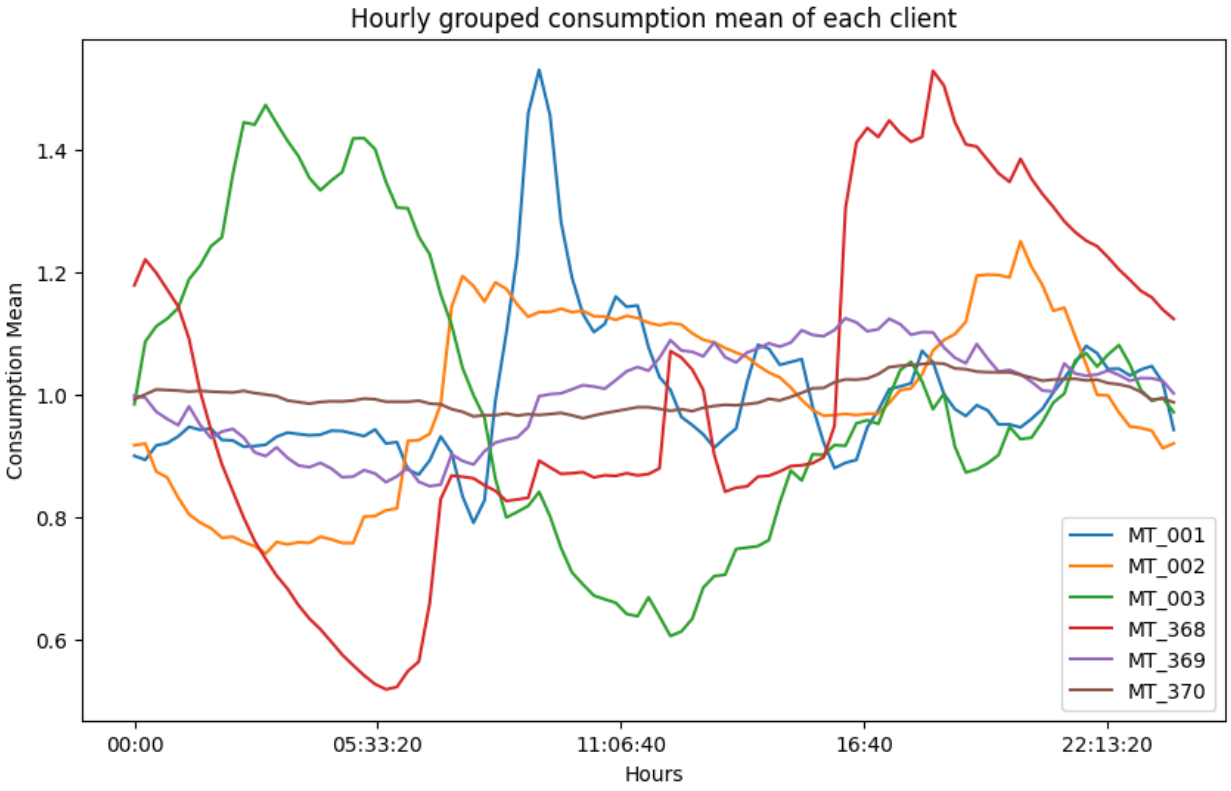


Εικόνα 10: Ετήσια μέση κατανάλωση κάθε πελάτη (στο γράφημα φαίνονται οι πρώτοι πέντε πελάτες για λόγους ευκρίνειας)

- Ωριαία μέση κατανάλωση των πελατών (σε 24 ώρες)

Time	MT_001	MT_002	MT_003	MT_004	MT_005	MT_006	MT_007	MT_008	MT_009	MT_010	...	MT_361	MT_362	MT_363	MT_364	MT_365	MT_366	MT_367	MT_368	MT_369	MT_370
00:00:00	4.763515	25.392998	3.829580	123.251217	59.036185	189.897788	7.095538	244.817395	52.181167	52.433090	...	114.240610	38080.839416	1457.093689	1940.817850	36.011001	17.294657	668.212448	148.853625	832.232512	11538.873545
00:15:00	4.728778	25.472168	4.228313	118.481470	56.584921	178.289016	6.822693	235.300081	50.546169	51.034063	...	110.032198	33987.773723	1100.857741	1736.168713	32.407760	16.812026	653.012311	154.201650	829.298623	11633.527323
00:30:00	4.853829	24.202843	4.325816	114.473918	54.372886	170.302724	6.623603	228.203569	49.258907	50.091241	...	103.861554	29096.350365	1056.615541	1618.011778	29.862056	16.005860	648.241190	151.432436	810.285201	11726.159006
00:45:00	4.873513	23.934831	4.374172	110.813156	52.816228	163.508538	6.482796	221.202291	47.800967	48.930618	...	90.995353	25733.120438	927.777264	1455.312707	28.959164	16.066189	641.314413	147.993005	800.456740	11715.673703
01:00:00	4.931407	23.026965	4.441552	105.507462	51.231752	155.275135	6.318779	215.048908	46.824754	48.114355	...	86.181924	21899.087591	899.735132	1419.542137	28.388165	16.437239	633.619419	144.515494	791.859762	11705.869008
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
22:45:00	5.459400	26.245055	4.082454	151.938312	68.491855	242.513360	7.988863	286.398044	62.090054	60.769367	...	216.400433	61836.587591	2445.101481	3522.146649	110.509712	17.391824	702.316829	149.976238	852.450660	11778.319195
23:00:00	5.510347	26.178214	3.920742	145.252878	67.374711	234.415733	7.836193	278.957458	60.526836	59.434110	...	186.528653	56282.481752	2320.998491	3376.658925	104.836599	17.154780	699.465852	147.699024	856.231404	11679.029394
23:15:00	5.541610	26.051671	3.851776	138.257744	65.631120	222.529979	7.657219	270.610607	57.862986	57.076564	...	163.585578	51019.799270	2163.517509	2965.245521	83.515736	17.330961	690.365797	146.355544	855.907646	11530.144013
23:30:00	5.400348	25.263210	3.869216	133.495416	63.353436	210.633364	7.542717	262.604758	55.286866	55.317479	...	142.206818	44959.489051	1941.436447	2478.205873	56.415887	16.822703	686.346071	143.822429	852.245970	11556.776485
23:45:00	4.989301	25.478658	3.778847	128.404842	61.337235	200.577859	7.289470	253.394652	53.130264	53.635900	...	114.289454	41001.459854	1567.252610	2029.425182	40.834753	17.232193	678.009427	141.907742	835.219915	11481.012034

Εικόνα 11: Ωριαία μέση κατανάλωση κάθε πελάτη



Εικόνα 12: Ωριαία μέση κατανάλωση κάθε πελάτη (στο γράφημα φαίνονται οι πρώτοι τρείς και οι τελευταίοι τρεις πελάτες για λόγους ευκρίνειας)

1.3. Μετατροπή και κλιμάκωση των δεδομένων

Το τελευταίο βήμα στην διαδικασία εξοικείωσης με τα δεδομένα είναι η μετατροπή τους από KW σε KWH και στη συνέχεια η κλιμάκωση τους. Οι τεχνικές κλιμάκωσης δεδομένων (scaling) σχετίζονται μόνο με τα αριθμητικά χαρακτηριστικά και ανάγονται στην αναπαράστασή τους στην ίδια κλίμακα. Για το παρόν σύνολο δεδομένων οι τιμές των χαρακτηριστικών θα αναπαρασταθούν στο διάστημα [0,1]. Η διαδικασία που ακολουθήθηκε είναι η εξής:

1. Μετατροπή των αριθμητικών χαρακτηριστικών από KW σε KWH.
2. Κλιμάκωση των αριθμητικών χαρακτηριστικών.

Στις εικόνες που ακολουθούν φαίνεται, ανά βήμα, η διαδικασία που περιγράφηκε παραπάνω (Εικόνες 13, 14 & 15)

	MT_001	MT_002	MT_003	MT_004	MT_005	MT_006	MT_007	MT_008	MT_009	MT_010	...	MT_361	MT_362	MT_363	MT_364	MT_365	MT_366	MT_367	MT_368	MT_369	MT_370
0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0
1	3.807107	22.759602	77.324066	136.178862	70.731707	351.190476	9.609949	279.461279	75.174825	87.096774	...	128.479657	28500.0	1729.957806	1704.545455	15.645372	12.873025	504.828797	63.439065	761.730205	0.0
2	5.076142	22.759602	77.324066	136.178862	73.170732	354.166667	9.044658	279.461279	73.426573	84.946237	...	127.765882	26400.0	1654.008439	1659.090909	15.645372	13.458163	525.021949	60.100167	702.346041	0.0
3	3.807107	22.759602	77.324066	140.243902	69.512195	348.214286	8.479367	279.461279	75.174825	91.397849	...	114.204140	25200.0	1333.333333	1636.363636	15.645372	10.532475	526.777875	56.761269	696.480938	0.0
4	3.807107	22.759602	77.324066	140.243902	75.609756	339.285714	7.348785	279.461279	68.181818	88.172043	...	112.062812	23800.0	1324.894515	1636.363636	15.645372	14.628438	539.947322	63.439065	693.548387	0.0

Μετατροπή KW σε KWH

	MT_001	MT_002	MT_003	MT_004	MT_005	MT_006	MT_007	MT_008	MT_009	MT_010	...	MT_361	MT_362	MT_363	MT_364	MT_365	MT_366	MT_367	MT_368	MT_369	MT_370
0	0.000000	0.0000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00000	0.000000	0.000000	...	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0
1	0.951777	5.6899	19.331017	34.044715	17.682927	87.797619	2.402487	69.86532	18.793706	21.774194	...	32.119914	7125.0	432.489451	426.136364	3.911343	3.218256	126.207199	15.859766	190.432551	0.0
2	1.269036	5.6899	19.331017	34.044715	18.292683	88.541667	2.261164	69.86532	18.356643	21.236559	...	31.941470	6600.0	413.502110	414.772727	3.911343	3.364541	131.255487	15.025042	175.586510	0.0
3	0.951777	5.6899	19.331017	35.060976	17.378049	87.053571	2.119842	69.86532	18.793706	22.849462	...	28.551035	6300.0	333.333333	409.090909	3.911343	2.633119	131.694469	14.190317	174.120235	0.0
4	0.951777	5.6899	19.331017	35.060976	18.902439	84.821429	1.837196	69.86532	17.045455	22.043011	...	28.015703	5950.0	331.223629	409.090909	3.911343	3.657109	134.986831	15.859766	173.387097	0.0

Κλιμάκωση στο διάστημα [0,1]

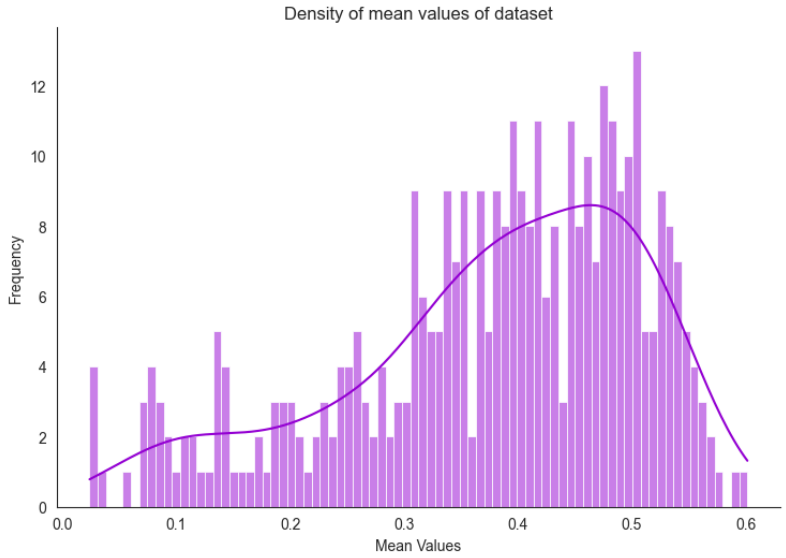
	MT_001	MT_002	MT_003	MT_004	MT_005	MT_006	MT_007	MT_008	MT_009	MT_010	...	MT_361	MT_362	MT_363	MT_364	MT_365	MT_366	MT_367	MT_368	MT_369	MT_370
0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0
1	0.078947	0.197531	0.511494	0.424051	0.471545	0.655556	0.215190	0.506098	0.477778	0.437838	...	0.150628	0.147822	0.223190	0.137615	0.046693	0.213592	0.443331	0.175115	0.491718	0.0
2	0.105263	0.197531	0.511494	0.424051	0.487805	0.661111	0.202532	0.506098	0.466667	0.427027	...	0.149791	0.136929	0.213391	0.133945	0.046693	0.223301	0.461064	0.165899	0.453384	0.0
3	0.078947	0.197531	0.511494	0.436709	0.463415	0.650000	0.189873	0.506098	0.477778	0.459459	...	0.133891	0.130705	0.172020	0.132110	0.046693	0.174757	0.462606	0.156682	0.449598	0.0
4	0.078947	0.197531	0.511494	0.436709	0.504065	0.633333	0.164557	0.506098	0.433333	0.443243	...	0.131381	0.123444	0.170931	0.132110	0.046693	0.242718	0.474171	0.175115	0.447705	0.0

Εικόνα 13: Μετατροπή και κλιμάκωση των δεδομένων

1.4. Απαντήσεις ερωτήματος 1

- Ποια είναι η κατανομή των τιμών κατανάλωσης στο συνόλο δεδομένων; Ποια είναι η ελάχιστη/μέση/μέγιστη τιμή ενέργειας ανά πελάτη; Παρατηρείτε φαινόμενα trend/seasonality στις χρονοσειρές, και αν ναι πού παρατηρούνται;

Αρχικά, κάναμε μια πρόβλεψη σχετικά με την κατανομή των δεδομένων. Συγκεκριμένα φτιάξαμε ένα distribution plot με την μέση τιμή κάθε στήλης των κλιμακωμένων δεδομένων και υποθέσαμε ότι το σύνολο δεδομένων θα ακολουθεί κανονική κατανομή σύμφωνα με την Εικόνα 14.



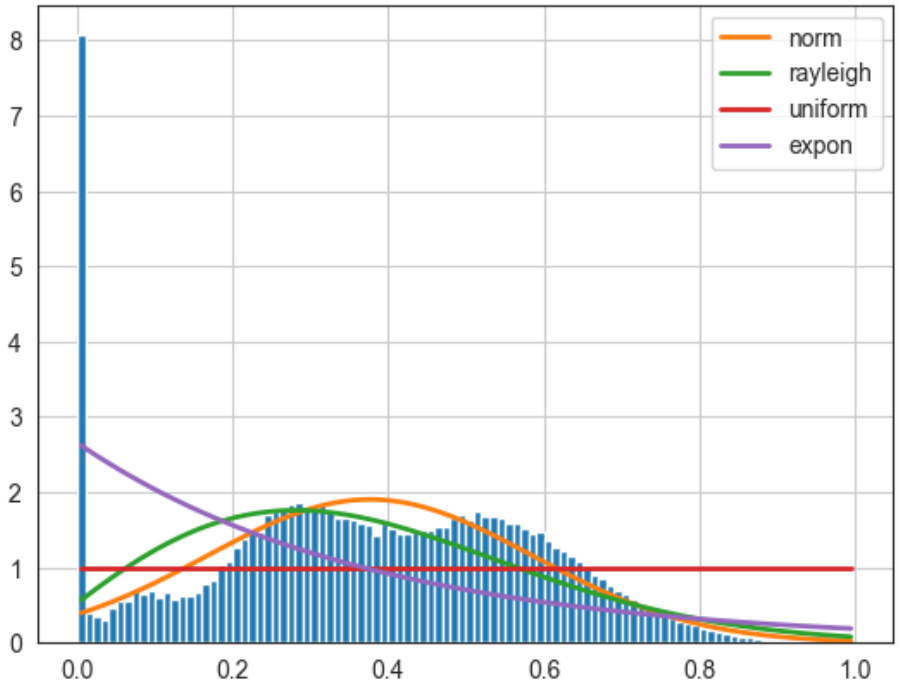
Εικόνα 14: Distribution plot με τη μέση τιμή κάθε πελάτη από τα κλιμακωμένα δεδομένα

Για να επαληθεύσουμε την υπόθεση μας χρησιμοποιήσαμε ένα instance της Fitter class, το οποίο δοκίμασε τις πιο κοινές κατανομές στα δεδομένα μας (Εικόνες 14 & 15). Τα αποτελέσματα ήταν αναμενόμενα, καθώς η κανονική κατανομή ταίριαζε

καλύτερα στα δεδομένα (Εικόνα 15 – πορτοκαλί).

	sumsquare_error	aic	bic	kl_div	ks_statistic	ks_pvalue
norm	63.012297	99.786809	-5.190893e+08	0.091326	0.042353	0.0
rayleigh	68.570538	67.321985	-5.157985e+08	0.153605	0.066482	0.0
uniform	90.678018	4.000000	-5.049194e+08	0.731198	0.256318	0.0
expon	95.275666	74.245991	-5.029940e+08	0.437662	0.213850	0.0
chi2	inf	inf	inf	inf	NaN	NaN

Εικόνα 14: Αποτελέσματα fitter instance σε πίνακα.



Εικόνα 15: Αποτελέσματα fitter instance σε γράφημα (με τα διαφορετικά χρώματα προσδιορίζονται οι κατανομές που δοκιμάστηκαν).

Όσον αφορά την μέγιστη και την ελάχιστη τιμή ενέργειας ανά πελάτη, τα αποτελέσματα φαίνονται στην Εικόνα 3. Τέλος, σύμφωνα με την Εικόνα 16, όπου φαίνονται οι μηνιαίες χρονοσειρές των πρώτων 10 πελατών, παρατηρούμε κάποια φαινόμενα trend στις χρονοσειρές, καθώς υπάρχουν αρκετά απότομες αυξομειώσεις στις τιμές τους. Για παράδειγμα στον πελάτη "MT_007" υπάρχει άνοδος κατανάλωσης κατά το χρονικό διάστημα Ιούνιος – Οκτώβριος, γεγονός το οποίο δίνει την δυνατότητα να κάνουμε διάφορες υποθέσεις για την κατηγορία του.

Εικόνα 16: Μηνιαίες χρονοσειρές των πρώτων δέκα πελατών

- Ποια είναι - με βάση τα δεδομένα - η δειγματοληψία των χρονικών ακολουθιών; Υπάρχουν κενά στη δειγματοληψία ή/και ελλείπεις τιμές;

  Η δειγματοληψία των χρονικών ακολουθιών έγινε κάθε 15 λεπτά από την 1η Ιανουαρίου 2011 έως και την 1η Ιανουαρίου 2015. Παρατηρήθηκε πως υπήρχαν κάποια κενά στην δειγματοληψία, όπως για παράδειγμα τα δείγματα που αντιστοιχούν στο 2011, τα οποία ήταν μηδενικές τιμές και διαγράφηκαν (βλ. 1.1). Τέλος, δεν υπήρχαν null τιμές στο σύνολο δεδομένων (Εικόνα 4).

- Ποια είναι η μονάδα μέτρησης της καταγεγραμμένης ενέργειας; Μετατρέψτε την σε KWH.

  Η μονάδα μέτρησης της καταγεγραμμένης ενέργειας είναι σε KW. Για την μετατροπής σε KWH, απαιτείται η διαίρεση των τιμών με το 4 (Εικόνα 13).

2. Μέρος ΄Β – Συσταδοποίηση

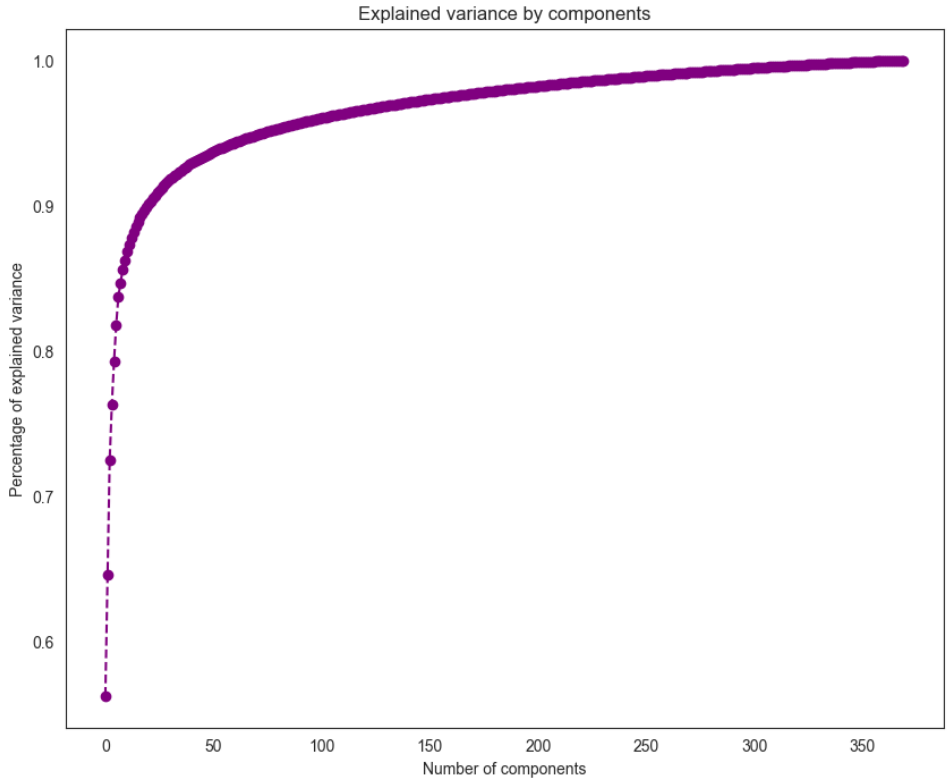
Η ομαδοποίηση ή συσταδοποίηση (αγγλικά: clustering), αποτελεί τομέας της μηχανικής μάθησης χωρίς επίβλεψη και της εξόρυξης δεδομένων. Πρόκειται για την διαδικασία κατά την οποία ένας αλγόριθμος χωρίζει ένα δοθέν σύνολο δεδομένων σε ομάδες ομοειδών αντικειμένων. Αντικειμενικός στόχος στην συσταδοποίηση είναι το να δημιουργούνται ομάδες που διαχωρίζουν όσο το δυνατόν γίνεται πιο ορθά τα δεδομένα. Για να είναι επιτυχημένη μια τεχνική ομαδοποίησης, πρέπει τα στοιχεία μιας συστάδας να μοιάζουν όσο γίνεται περισσότερο ενώ τα στοιχεία διαφορετικών συστάδων να διαφέρουν όσο γίνεται περισσότερο. Στην παρούσα εργασία θα εφαρμοστούν οι παρακάτω αλγόριθμοι ομαδοποίησης, με τη χρήση του πακέτου Scikit Learn:

- K-means
- DBSCAN

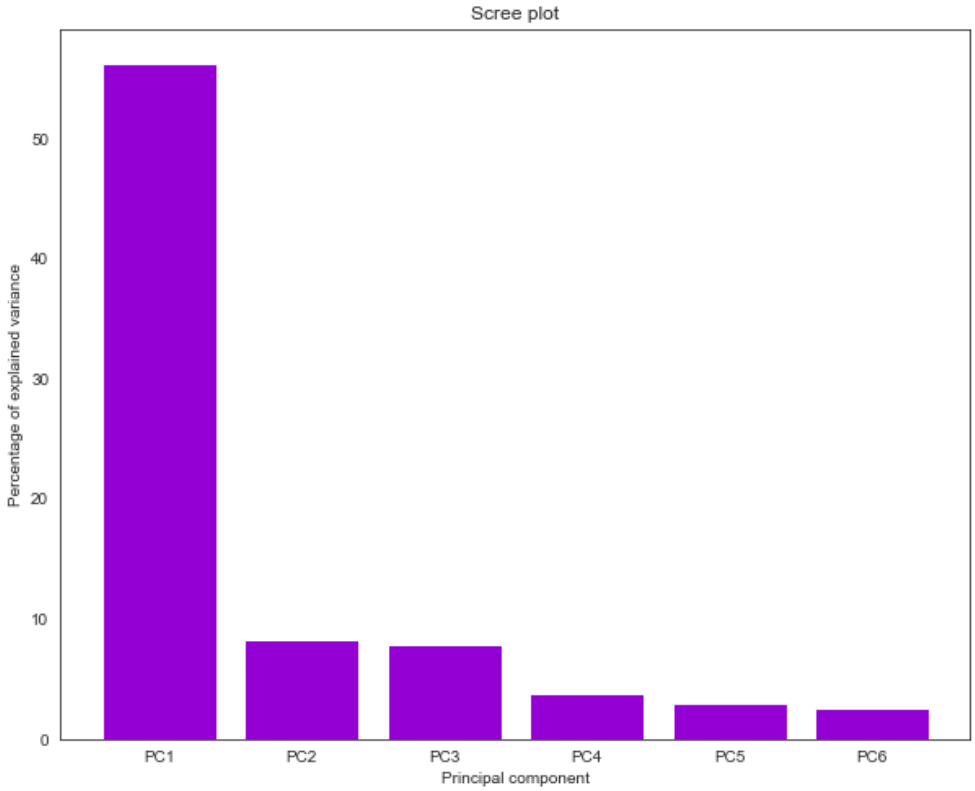
2.1. K-means

Αρχικά, λόγω του μεγάλου πλήθους (ως προς τη διάσταση του χρόνου) εγγραφών διανύσματος, χρησιμοποιήσαμε έναν αλγόριθμο μείωσης διαστάσεων (PCA) προκειμένου να διευκολυνθεί τόσο η διαδικασία υπολογισμού αποστάσεων όσο και η οπτικοποίηση του τελικού αποτελέσματος. Συγκεκριμένα, επειδή ο PCA λειτουργεί για την μείωση των στηλών, ενώ εμείς θέλουμε να μειώσουμε τις γραμμές (εγγραφές) κάναμε fit τον ανάστροφο πίνακα των κλιμακωμένων δεδομένων. Στη συνέχεια, είναι απαραίτητο να αποφασισθεί ο αριθμός των components που θα διατηρηθούν με βάση τα παρακάτω γραφήματα αθροιστικής διακύμανσης (Εικόνες 17 & 18), τα οποία απεικονίζουν το ποσοστό διακύμανσης ανάλογα με τα components που υπάρχουν στον άξονα x. Ένας εμπειρικός κανόνας είναι να διατηρείται περίπου το 80% της διακύμανσης, οπότε αποφασίσαμε να διατηρήσουμε έξι χαρακτηριστικά. Επιπλέον είναι ολοφάνερο πως τα δύο πρώτα δύο components φαίνεται να είναι αυτά που μας ενδιαφέρουν περισσότερο. Τέλος, εφαρμόζουμε πάλι τον αλγόριθμο PCA για έξι components και μειώνεται το πλήθος των εγγραφών του συνόλου δεδομένων (διάσταση χρόνου). Στην Εικόνα 19 φαίνονται τα αποτελέσματα του PCA μαζί με τα κλιμακωμένα δεδομένα σε ανάστροφη μορφή.

Το επόμενο βήμα είναι να εφαρμοσθεί ο αλγόριθμος K-means με βάση τα αποτελέσματα που έδωσε ο PCA και επιλέγοντας οκτώ clusters, όπως υποδεικνύεται και από την εκφώνηση. Ο K-means κατηγοριοποιεί τα δεδομένα σε οκτώ clusters (labeling) (Εικόνα 20) και, τέλος, οπτικοποιούμε τα αποτελέσματα σε σχέση με τα δύο πρώτα components σε ένα scatter plot (Εικόνα 21) και οι ξεχωριστές συστάδες είναι πολύ ευδιάκριτες.



Εικόνα 17: Αθροιστική διακύμανση ανά component



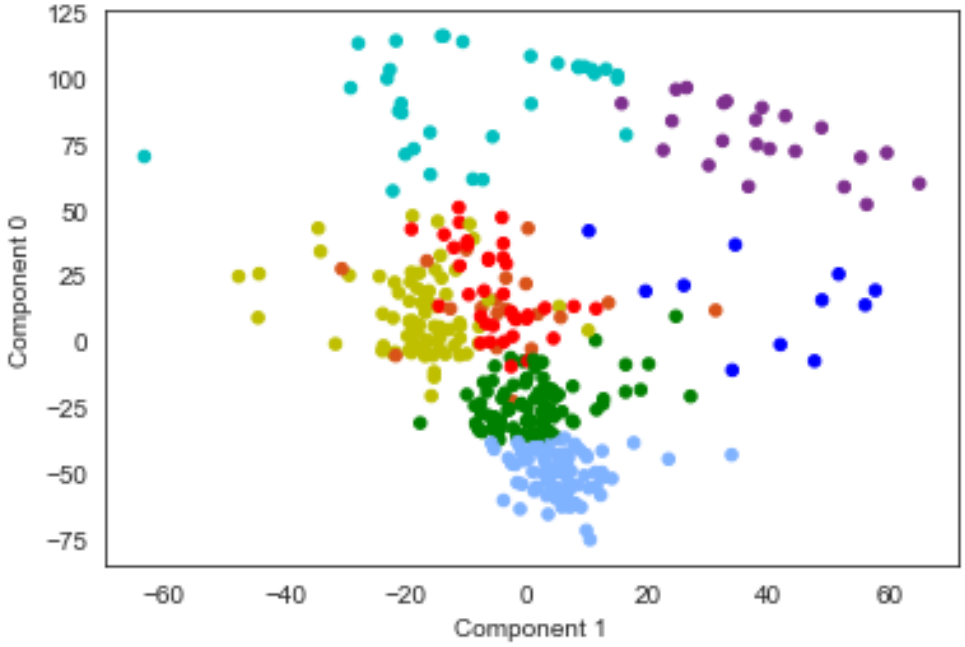
Εικόνα 18: Αθροιστική διακύμανση ανά component

2	3	4	5	6	7	8	9	...	105212	105213	105214	105215	Component 0	Component 1	Component 2	Component 3	Component 4	Component 5
105263	0.078947	0.078947	0.105263	0.078947	0.078947	0.131579	0.078947	...	0.052632	0.052632	0.052632	0.026316	86.619111	-20.749947	-10.574293	10.813827	-7.510341	-12.277141
197531	0.197531	0.197531	0.191358	0.197531	0.197531	0.216049	0.203704	...	0.191358	0.185185	0.179012	0.185185	44.976418	-11.204342	-1.634519	4.107970	-7.615981	-14.280154
511494	0.511494	0.511494	0.511494	0.511494	0.511494	0.511494	0.511494	...	0.011494	0.011494	0.011494	0.011494	113.118555	-27.929702	-14.732696	9.768825	-6.503747	-11.266170
424051	0.436709	0.436709	0.455696	0.417722	0.411392	0.424051	0.417722	...	0.468354	0.518987	0.506329	0.518987	17.600291	-13.419135	22.559565	-17.976715	2.557194	4.860085
487805	0.463415	0.504065	0.487805	0.487805	0.447154	0.447154	0.455285	...	0.569106	0.544715	0.552846	0.569106	22.308462	-19.297394	27.275858	-8.077855	4.443689	2.703872

Εικόνα 19: Αποτελέσματα PCA (component 0 – 5) και κλιμακωμένα δεδομένα σε ανάστροφη μορφή

```
0       third
1      second
2       third
3       fifth
4       fifth
        ...
365     third
366     first
367    eighth
368     sixth
369    fourth
Name: Segment, Length: 370, dtype: object
```

Εικόνα 20: Κατηγοριοποίηση των 370 πελατών σε οκτώ clusters από τον K-means



Εικόνα 21: Αποτελέσματα K-means σε σχέση με τα δύο πρώτα components

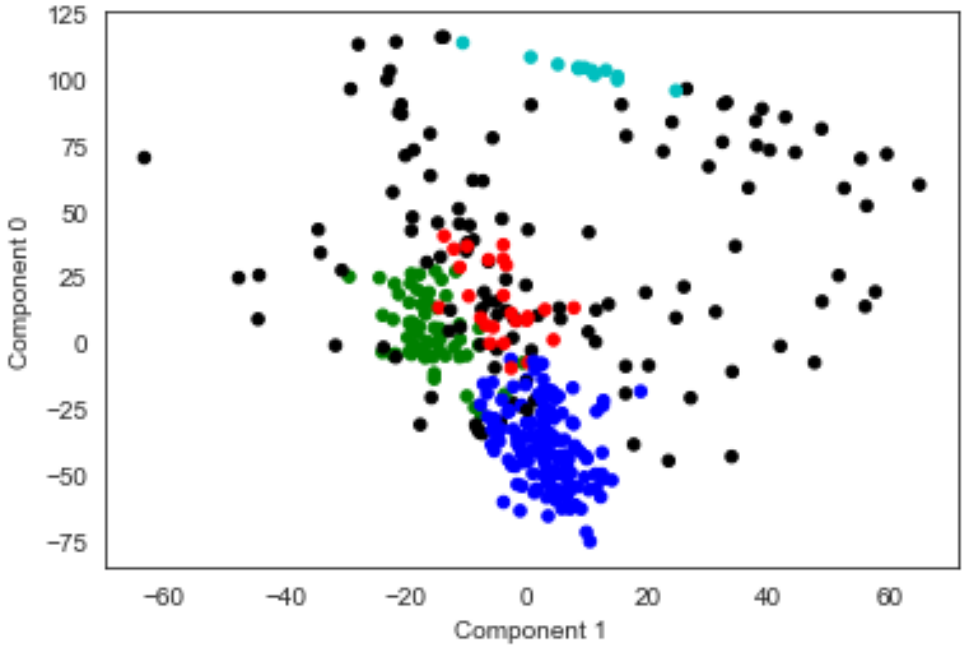
Έστω ότι δεν είχαμε την πληροφορία του domain expert. Πώς μπορούμε να οδηγηθούμε στη βέλτιστη τιμή της παραμέτρου k;

Στην περίπτωση που δεν υπήρχαν υποδείξεις από τον domain expert και χρειαζόταν να προσδιορίσουμε εμείς των αριθμό των clusters θα εφαρμόζαμε μια πολύ γνωστή προσέγγιση για τον προσδιορισμό των συστάδων, την μέθοδο Elbow. Συγκεκριμένα, θα δοκιμάζαμε τον K-means με διαφορετικές τιμές στην παράμετρο k και για κάθε δοκιμή θα υπολογίζαμε το άθροισμα τετραγώνων. Στη συνέχεια θα

δημιουργούσαμε ένα καμπυλοειδές γράφημα με τα αθροίσματα τετραγώνων για κάθε τιμή που δοκιμάστηκε στην παράμετρο k. Ο ιδανικός αριθμός συστάδων είναι το σημείο στο οποίο «λυγίζει» (κάνει elbow) η καμπύλη.

2.2. DBSCAN

Αρχικά, σύμφωνα με την παράγραφο 2.1, αξίζει να σημειωθεί πως ο αλγόριθμος DBSCAN θα εφαρμοσθεί με βάση τα αποτελέσματα που έδωσε ο PCA. Όσον αφορά τις τιμές των παραμέτρων min_samples και eps, μετά από δοκιμές του DBSCAN για διάφορες τιμές των εν λόγω, παραμέτρων καταλήξαμε στο συμπέρασμα πως ο αλγόριθμος δίνει καλύτερα αποτελέσματα για eps=16,5 και min_samples=7. Οπότε, με βάση τα παραπάνω εφαρμόσθηκε ο αλγόριθμος και κατηγοριοποίησε τα δεδομένα σε πέντε συστάδες, συμπεριλαμβανομένου του θορύβου, όπως φαίνεται και στην Εικόνα 22 (η οπτικοποίηση των αποτελεσμάτων του DBSCAN γίνεται και πάλι σε σχέση με τα δύο πρώτα components του PCA).



Εικόνα 22: Αποτελέσματα DBSCAN σε σχέση με τα δύο πρώτα components

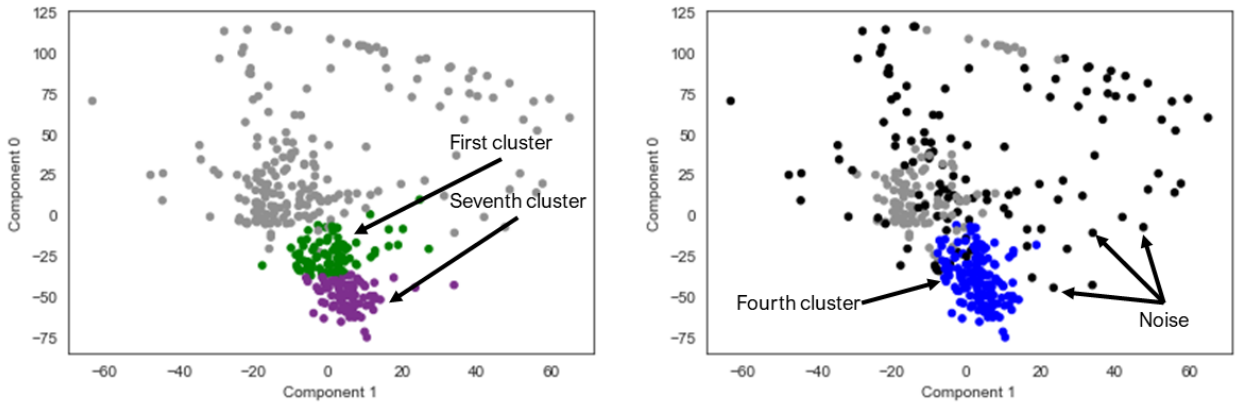
2.3. K-means VS DBSCAN

Μετά την εφαρμογή των δύο αλγόριθμων ομαδοποίησης, παρατηρούμε κάποιες αισθητές διαφορές, όσον αφορά τα αποτελέσματα τους σχετικά με το παρόν σύνολο δεδομένων. Αρχικά, συγκρίνοντας τα δύο scatter plots, παρατηρούμε πως ο DBSCAN φαίνεται να έχει κάνει καλύτερη κατηγοριοποίηση στα δεδομένα, σε σχέση με τον K-means, καθώς οι συστάδες που έχει δημιουργήσει φαίνεται να είναι πιο πυκνές και ομοιογενείς. Επιπλέον, τα σημεία που έχουν ανιχνευθεί ως θόρυβος στον DBSCAN, παρουσιάζονται ως συστάδες στον K-means, αλλά δεν έχουν την μορφή πυκνών συστάδων, γεγονός το οποίο μας κάνει να αμφιβάλουμε για την αποτελεσματικότητα του K-means. Οι διαφορές γίνονται ακόμα

πιο αισθητές αν κοιτάξουμε τον παρακάτω πίνακα (Εικόνα 23), στον οποίον απεικονίζεται η κατηγοριοποίηση που έχει κάνει ο αλγόριθμος K-means για την τέταρτη συστάδα του αλγορίθμου DBSCAN (μπλε – Εικόνα 22). Παρατηρούμε πως η ομαδοποίηση του K-means είναι πολύ διαφορετική σε σχέση με αυτή του DBSCAN (Εικόνα 24).

Segment K-means PCA	Segment	Segment DBSCAN PCA	DBSCAN Segment
0	first	3	fourth
0	first	3	fourth
0	first	3	fourth
0	first	3	fourth
0	first	3	fourth
6	seventh	3	fourth
6	seventh	3	fourth
6	seventh	3	fourth
6	seventh	3	fourth
0	first	3	fourth
0	first	3	fourth
6	seventh	3	fourth
6	seventh	3	fourth
6	seventh	3	fourth
6	seventh	3	fourth

Εικόνα 23: Ομαδοποίηση του K-means για το τέταρτο cluster του DBSCAN



Εικόνα 24: Ομαδοποίηση K-means VS DBSCAN

3. Μέρος ΄Γ – Πρόβλεψη

Εφόσον έχουν προηγηθεί οι διαδικασίες προετοιμασίας και ομαδοποίησης των δεδομένων, το επόμενο βήμα είναι να χρησιμοποιήσουμε μοντέλα Μηχανικής Μάθησης, με σκοπό την πρόβλεψη της κατανάλωσης ενέργειας τόσο βραχυπρόθεσμα (εντός τριών ωρών) όσο και μακροπρόθεσμα (εντός τριών ημερών).

3.1. Μακροπρόθεσμη Πρόβλεψη

3.1.1. Κλασσικοί Αλγόριθμοι Μηχανικής Μάθησης

Αρχικά, επιλέγουμε να δουλέψουμε με την τέταρτή συστάδα που δημιούργησε ο αλγόριθμος DBSCAN. Με βάση αυτή τη συστάδα θα προβλέψουμε πρώτον, την συνολική ημερήσια μακροπρόθεσμη κατανάλωση κάθε πελάτη τις επόμενες τρείς μέρες και στη συνέχεια την αντίστοιχη μακροπρόθεσμη κατανάλωση, ανά δεκαπέντε λεπτά. Από τον πίνακα της Εικόνας 23, διαγράφουμε τα components και τα segments από την διαδικασία της ομαδοποίησης, ώστε να κρατήσουμε μόνο τα δεδομένα μέσα από τα οποία θα γίνει η πρόβλεψη. Επιπλέον, επειδή για τις ανάγκες χρήσης του PCA ο συγκεκριμένος πίνακας ήταν σε ανάστροφη μορφή, κρίθηκε απαραίτητο να μετατραπεί στην μορφή που ήταν πριν, δηλαδή η κάθε στήλη να αντιπροσωπεύει και από έναν πελάτη (Εικόνα 25).

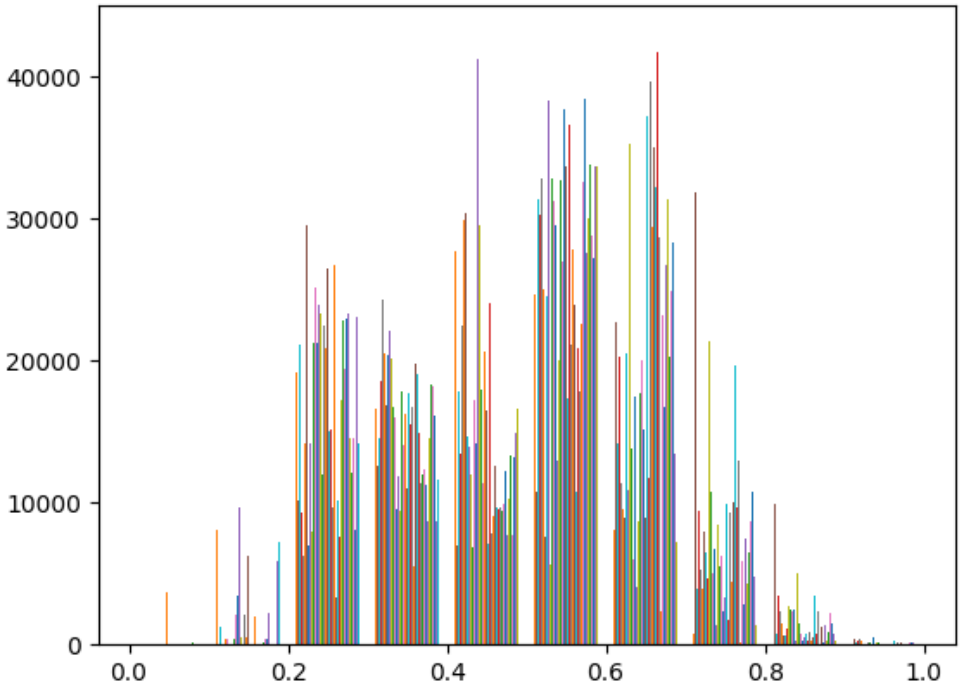
	90	125	134	136	138	165	167	168	170	171	...	324	325	326	327	328	329	330	344	352	368
0	0.000000	0.000000	0.000000	0.000000	0.000000	0.355646	0.240324	0.309800	0.417311	0.333077	...	0.238182	0.320538	0.240963	0.280899	0.343174	0.175739	0.351351	0.000000	0.000000	0.000000
1	0.354348	0.176136	0.298611	0.208955	0.577869	0.302551	0.211794	0.286806	0.410019	0.308976	...	0.230909	0.333148	0.243973	0.269663	0.327369	0.159249	0.373464	0.156660	0.196114	0.491718
2	0.306522	0.153409	0.298611	0.208955	0.577869	0.260118	0.232184	0.272113	0.417311	0.304118	...	0.223636	0.290247	0.249992	0.258427	0.327369	0.167481	0.351351	0.154784	0.271403	0.453384
3	0.500000	0.153409	0.215278	0.273632	0.616803	0.244152	0.224004	0.264766	0.410019	0.318597	...	0.225455	0.330626	0.243973	0.247191	0.351056	0.181227	0.353808	0.156660	0.307225	0.449598
4	0.510870	0.164773	0.243056	0.213930	0.615779	0.228240	0.219934	0.302454	0.424603	0.318597	...	0.229091	0.330626	0.256012	0.255618	0.347115	0.153762	0.351351	0.151032	0.305404	0.447705

5 rows × 151 columns

Εικόνα 25: Δεδομένα κάθε πελάτη από την τέταρτη συστάδα του DBSCAN

Στη συνέχεια, είναι απαραίτητο να αναλύσουμε τη συμπεριφορά των χρονοσειρών που ανήκουν στην, εν λόγω συστάδα και να τις προετοιμάσουμε για την είσοδό τους στο μοντέλο μηχανικής μάθησης. Όπως φαίνεται και στις παρακάτω εικόνες, τα δεδομένα φαίνεται να ακολουθούν κανονική κατανομή (Εικόνα 26) και οι τιμές τους είναι στην κλίμακα που ορίσαμε κατά την διαδικασία της κλιμάκωσης, κοιτάζοντας την μέγιστη και την ελάχιστη τιμή κάθε πελάτη (Εικόνα 27). Ωστόσο παρατηρούμε πως η μέγιστη τιμή του συγκεκριμένου data frame είναι μεγαλύτερη από το 1 (1.0000000000000002), επομένως για τεχνικούς λόγους βεβαιώνουμε πως όλες οι τιμές βρίσκονται στο διάστημα [0,1], εφαρμόζοντας την παρακάτω μέθοδο αντικατάστασης:

- Για κάθε τιμή x όπου x > 1 => x = 1
- Για κάθε τιμή x όπου x < 0 => x = 0

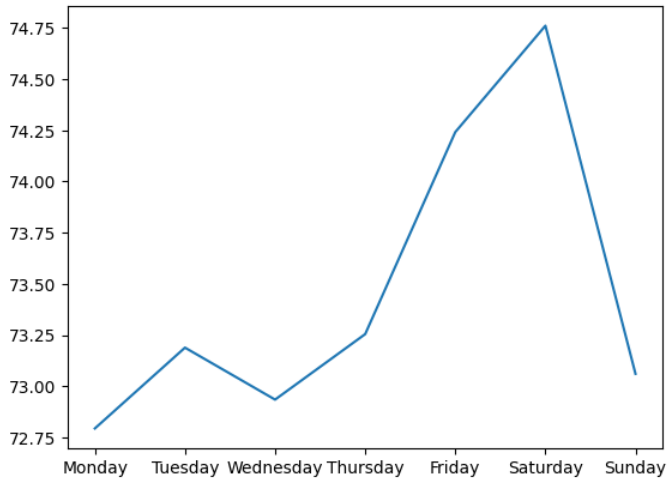


Εικόνα 26: Κατανομή των δεδομένων της επιλεγμένης συστάδας

	90	125	134	136	138	165	167	168	170	171	...	324	325	326
ount	105216.000000	105216.000000	105216.000000	105216.000000	105216.000000	105216.000000	105216.000000	105216.000000	105216.000000	105216.000000	...	105216.000000	105216.000000	105216.000000
nean	0.450782	0.416856	0.434123	0.412891	0.479220	0.601471	0.482281	0.503433	0.575218	0.464101	...	0.506067	0.482046	0.502653
std	0.125066	0.141954	0.213096	0.176229	0.162813	0.183578	0.155667	0.142341	0.144582	0.151389	...	0.160018	0.163468	0.164780
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.000000
25%	0.358696	0.289773	0.201389	0.218905	0.314549	0.445871	0.346221	0.361446	0.431967	0.318597	...	0.410909	0.315493	0.346396
50%	0.443478	0.437500	0.506944	0.442786	0.517418	0.663502	0.530748	0.538422	0.610617	0.492399	...	0.494545	0.521022	0.537336
75%	0.526087	0.528409	0.618056	0.552239	0.608607	0.743118	0.600138	0.622612	0.691556	0.574466	...	0.609091	0.601957	0.624891
max	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	...	1.000000	1.000000	1.000000

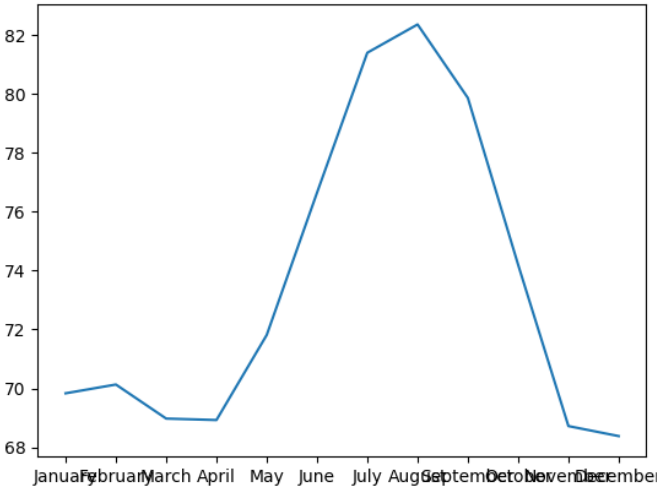
Εικόνα 27: Περιγραφικά στατιστικά των δεδομένων της επιλεγμένης συστάδας

Έπειτα, πρέπει να ελέγξουμε τα δεδομένα σε περίπτωση που υπάρχουν φαινόμενα seasonality σε ημερήσια (Εικόνες 28, 29), μηνιαία (Εικόνες 30, 1) και ετήσια βάση (Εικόνες 32, 33), υπολογίζοντας το αντίστοιχο άθροισμα για κάθε πελάτη. Τα συμπεράσματά μας, βάσει των παρακάτω εικόνων είναι πως σε ημερήσια και ετήσια βάση δεν παρατηρείται κάποιο φαινόμενο seasonality, σε σχέση με την μηνιαία βάση, όπου παρατηρούμε μια απότομη αύξηση τους θερινούς μήνες.




```
Day Name
Monday        72.794613
Tuesday       73.189144
Wednesday     72.935378
Thursday      73.255146
Friday        74.241470
Saturday      74.760704
Sunday        73.061335
dtype: float64
```

Εικόνες 28, 29: Ημερήσια αθροιστική χρονοσειρά των πελατών της επιλεγμένης συστάδας




```
Month Name
January       69.834597
February      70.132825
March         68.976412
April         68.927080
May           71.815696
June          76.645596
July          81.397091
August        82.358398
September     79.863048
October       74.199468
November      68.721957
December      68.381716
dtype: float64
```

Εικόνες 30,31: Μηνιαία αθροιστική χρονοσειρά των πελατών της επιλεγμένης συστάδας


```
Year
2012    75.889737
2013    73.105163
2014    71.380952
dtype: float64
```

Εικόνες 32, 33: Ετήσια αθροιστική χρονοσειρά των πελατών της επιλεγμένης συστάδας

Καθώς οι χρονοσειρές του συνόλου δεδομένων είναι λίγες στο πλήθος, θα ακολουθήσουμε την προσέγγιση επαύξησης των δεδομένων με την κατάτμησή τους μέσω ημι-επικαλυπτόμενων παραθύρων. Για τις ανάγκες αυτής της προσέγγισης δημιουργήσαμε δύο συναρτήσεις;

- divide_chunks(): όπου δέχεται ως παραμέτρους μια χρονοσειρά και το window και χωρίζει την χρονοσειρά σε segments.

- get_labels(): όπου δέχεται έναν πίνακα με τα segments μιας χρονοσειράς και βρίσκει την συνολική κατανάλωση των επόμενων τριών ημερών, την οποία ορίζει ως label του συγκεκριμένου segment.

Επομένως, καλούμε τις δύο παραπάνω συναρτήσεις για κάθε καταναλωτή της επιλεγμένης συστάδας με μέγεθος segment 16 ημέρες και μέγεθος label 3 ημέρες. Επιπλέον, αξίζει να σημειωθεί πως, σύμφωνα με τον αριθμό δεδομένων που έχουμε κρίναμε πως η καλύτερη τιμή για το window ήταν το 8, καθώς είναι το μισό μέγεθος κάθε segment και με αυτό τον τρόπο οι χρονοσειρές χωρίζονται ακριβώς. Τα αποτελέσματα αυτής της διαδικασίας φαίνονται στην παρακάτω εικόνα (Εικόνα 34). Ωστόσο, παρατηρούμε πως το τελευταίο segment περιέχει κάποιες NaN τιμές, γεγονός, το οποίο είναι αναμενόμενο λόγω της διαδικασίας που ακολουθήσαμε, οπότε τις αφαιρούμε από το dataframe των αποτελεσματων (Εικόνα 35).

0_segment_8	90_segment_9	...	368_segment_127	368_segment_128	368_segment_129	368_segment_130	368_segment_131	368_segment_132	368_segment_133	368_segment_134	368_segment_135	368_segment_136
0.354348	0.308696	...	0.525319	0.451491	0.476574	0.461903	0.447705	0.419309	0.464742	0.467582	0.421202	NaN
0.384783	0.319565	...	0.519167	0.456223	0.463322	0.465689	0.443445	0.424988	0.470421	0.460956	0.420256	NaN
0.336957	0.369565	...	0.552769	0.464269	0.482726	0.477047	0.454330	0.430194	0.467582	0.483673	0.432087	NaN
0.393478	0.352174	...	0.536678	0.458116	0.478467	0.467108	0.455750	0.430194	0.472788	0.467108	0.432560	NaN
0.384783	0.308696	...	0.496451	0.451018	0.478467	0.452911	0.443445	0.421202	0.458116	0.451018	0.417416	NaN

Εικόνα 34: Αποτελέσματα κατάτμησης χρονοσειρών μέσω επικαλυπτόμενων παραθύρων

0_segment_8	90_segment_9	...	368_segment_126	368_segment_127	368_segment_128	368_segment_129	368_segment_130	368_segment_131	368_segment_132	368_segment_133	368_segment_134	368_segment_135
153.830435	138.817391	...	171.803597	159.601988	142.718410	133.576905	139.269285	137.883578	143.055372	148.420256	141.697586	0.000000
0.313043	0.332609	...	0.537624	0.434927	0.531472	0.496451	0.447231	0.478467	0.459063	0.442972	0.417889	0.461429
0.252174	0.326087	...	0.531472	0.409371	0.518221	0.485092	0.446285	0.471368	0.468055	0.457170	0.434453	0.460009
0.273913	0.306522	...	0.508282	0.395173	0.492191	0.495031	0.432087	0.470421	0.471841	0.451018	0.432560	0.454330
0.432609	0.528261	...	0.503549	0.393753	0.488405	0.496451	0.429721	0.461429	0.458590	0.440133	0.411737	0.440606

Εικόνα 35: Αποτελέσματα κατάτμησης χρονοσειρών μέσω επικαλυπτόμενων παραθύρων (χωρίς NaN τιμές)

Παρατηρούμε πως η πρώτη σειρά είναι τα labels και πως το τελευταίο στοιχείο της πρώτης σειράς είναι μηδενικό, καθώς πρόκειται για το label που θέλουμε να προβλέψουμε. Έτσι απομονώνουμε τις στήλες του segment 135 σε ένα ξεχωριστό dataframe, το X (Εικόνα 36), όπου πάνω σε αυτό θα εφαρμόσουμε το τελικό μας μοντέλο, ώστε να προβλέψουμε τα μηδενικά labels. Για να καταλήξουμε στο τελικό μοντέλο, όμως, θα πρέπει πρώτα να εκπαιδεύσουμε διάφορους αλγόριθμους μηχανικής μάθησης και να επιλέξουμε τον καλύτερο.

	90_segment_135	125_segment_135	134_segment_135	136_segment_135	138_segment_135	165_segment_135	167_segment_135	168_segment_135	170_segment_135	171_segment_135	...	324_segment_135	32
0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	
1	0.258696	0.420455	0.131944	0.278607	0.300205	0.456479	0.329942	0.390905	0.307496	0.221474	...	0.343636	
2	0.260870	0.352273	0.131944	0.228856	0.267418	0.498966	0.354361	0.368792	0.344101	0.211853	...	0.332727	
3	0.258696	0.318182	0.131944	0.223881	0.262295	0.419350	0.362501	0.376212	0.351393	0.255195	...	0.321818	
4	0.397826	0.318182	0.138889	0.228856	0.254098	0.398080	0.342151	0.309800	0.322153	0.264816	...	0.320000	
...	...	...	...	...	...	...	...	...	...	...	...	...	
1532	0.315217	0.142045	0.173611	0.378109	0.372951	0.265422	0.301412	0.346679	0.292839	0.269627	...	0.221818	
1533	0.330435	0.153409	0.138889	0.278607	0.360656	0.244152	0.264743	0.272113	0.329444	0.240764	...	0.214545	
1534	0.328261	0.130682	0.131944	0.223881	0.363730	0.302551	0.276993	0.295034	0.322153	0.235905	...	0.212727	
1535	0.334783	0.125000	0.138889	0.218905	0.368852	0.265422	0.256603	0.295034	0.300204	0.235905	...	0.203636	
1536	0.319565	0.164773	0.131944	0.213930	0.361680	0.286639	0.244394	0.295034	0.262797	0.211853	...	0.190909	

Εικόνα 36: Segment 135 (η πρώτη σειρά είναι τα labels που θέλουμε να προβλέψουμε)

Το επόμενο βήμα είναι να χωρίσουμε τα δεδομένα σε δεδομένα εκπαίδευσης (train sets), τα οποία καταλαμβάνουν το 80% του συνόλου και ελέγχου (test sets), τα οποία καταλαμβάνουν το υπόλοιπο 20% του συνόλου. Με τη σειρά τους τα δεδομένα εκπαίδευσης και ελέγχου θα πρέπει να διαχωριστούν σε ετικέτες και δεδομένα. Ένα παράδειγμα διαχωρισμού του συνόλου εκπαίδευσης φαίνεται στις Εικόνες 36 και 37.

	0
90_segment_0	146.697826
90_segment_1	150.808696
90_segment_2	163.108696
90_segment_3	158.654348
90_segment_4	150.256522

Εικόνα 37: Ετικέτες εκπαίδευσης (y_train)

	1	2	3	4	5	6	7	8	9	10	...	1527	1528	1529	1530	1531	1532	1533	1534	1535	153
90_segment_0	0.000000	0.354348	0.306522	0.500000	0.510870	0.493478	0.495652	0.500000	0.510870	0.482609	...	0.378261	0.386957	0.354348	0.332609	0.397826	0.354348	0.356522	0.380435	0.391304	0.3847
90_segment_1	0.336957	0.334783	0.369565	0.543478	0.547826	0.526087	0.547826	0.541304	0.554348	0.528261	...	0.310870	0.308696	0.343478	0.360870	0.404348	0.369565	0.391304	0.406522	0.356522	0.3913
90_segment_2	0.345652	0.384783	0.352174	0.528261	0.530435	0.517391	0.534783	0.536957	0.552174	0.532609	...	0.332609	0.354348	0.343478	0.369565	0.443478	0.430435	0.386957	0.419565	0.408696	0.3934
90_segment_3	0.378261	0.376087	0.354348	0.550000	0.571739	0.582609	0.545652	0.576087	0.582609	0.558696	...	0.352174	0.356522	0.367391	0.397826	0.380435	0.436957	0.389130	0.395652	0.408696	0.4260
90_segment_4	0.395652	0.410870	0.382609	0.573913	0.580435	0.591304	0.630435	0.604348	0.621739	0.586957	...	0.345652	0.326087	0.313043	0.354348	0.452174	0.432609	0.404348	0.419565	0.382609	0.4239

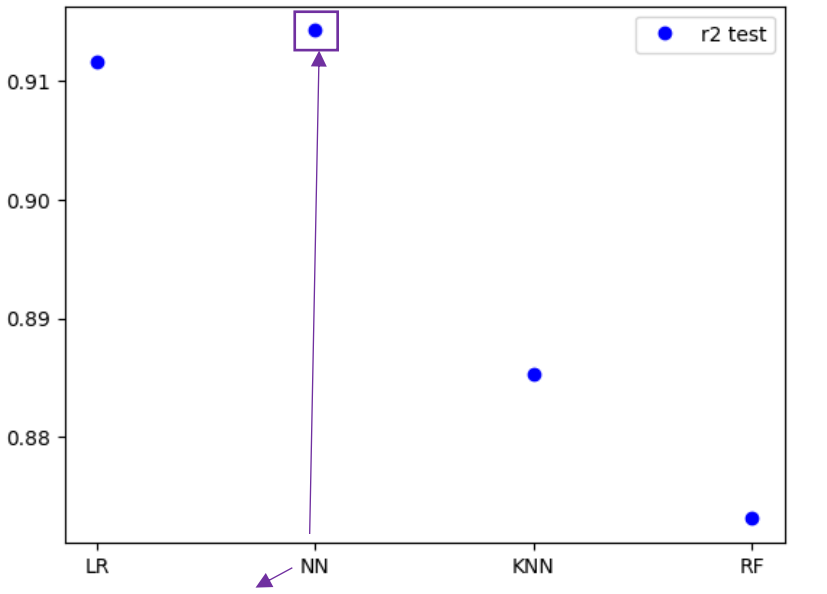
Εικόνα 38: Δεδομένα εκπαίδευσης (X_train)

Πλέον, μπορούμε να δημιουργήσουμε και να εκπαιδεύσουμε τα μοντέλα μηχανικής μάθησης, ώστε να επιλέξουμε το καλύτερο και να το χρησιμοποιήσουμε για την πρόβλεψή μας. Συγκεκριμένα υλοποιήσαμε τα παρακάτω μοντέλα μηχανικής μάθησης:

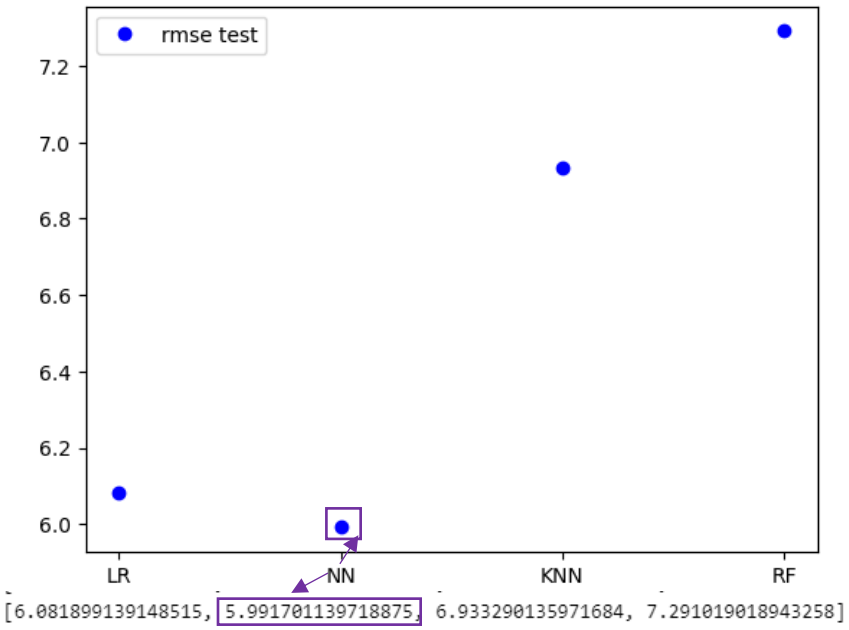
- Linear Regression (LR)
- Multi-Layer Perceptron Regression (MLPR)
- K-Nearest Neighbors Regression (KNN)
- Random Forest Regression (RF)

Συγκεκριμένα, έχουμε αποθηκεύσει κάποια χαρακτηριστικά των μοντέλων στον πίνακα models, όπως για παράδειγμα το όνομα τους και τις απαραίτητες παραμέτρους για το cross validation. Έπειτα για κάθε μοντέλο και τις αντίστοιχες παραμέτρους του, εφαρμόζεται μια μέθοδος GridSearchCV, όπου το αποτέλεσμα της εκπαιδεύεται με τα δεδομένα εκπαίδευσης. Τέλος, κάνουμε τις προβλέψεις με τα δεδομένα ελέγχου και για κάθε καλύτερο μοντέλο, υπολογίζεται το R-Squared και RMSE σκορ (Εικόνα

39). Με αυτόν τον τρόπο βρίσκουμε τις καλύτερες παραμέτρους και έτσι προκύπτει το καλύτερο μοντέλο από κάθε είδος (LR, MLPR, KNN, RF). Παρόλα αυτά, το τελικό μας μοντέλο, θα πρέπει να έχει, προφανώς, πολύ υψηλή R-Squared τιμή και πολύ χαμηλό RMSE. Επομένως κάνοντας κάποιες αριθμητικές και οπτικές συγκρίσεις στις παρακάτω εικόνες, καταλήγουμε στο συμπέρασμα πως το MLPR φαίνεται να είναι το πιο αξιόπιστο μοντέλο.



Εικόνα 39: R-Squared σκορ για κάθε καλύτερο μοντέλο



Εικόνα 40: RMSE για κάθε καλύτερο μοντέλο

Για να προβλέψουμε τη συνολική κατανάλωση για τις επόμενες τρείς μέρες θα ακολουθήσουμε την παρακάτω διαδικασία:

- Διαγράφουμε τα μηδενικά labels του dataframe X, καθώς αυτά θέλουμε να προβλέψουμε.
- Ορίζουμε ως train set, όλο το σύνολο δεδομένων και ως test set το X.
- Εφαρμόζουμε την μέθοδο GridSearchCV, ώστε να προκύψουν οι καλύτερες παράμετροι.
- Εκπαιδεύουμε το καλύτερο MLPR με το σύνολο δεδομένων εκπαίδευσης.
- Προβλέπουμε την συνολική κατανάλωση κάθε πελάτη για τις επόμενες τρείς μέρες με test set το X και το αποτέλεσμα που προκύπτει είναι τα labels του dataframe X (Εικόνα 41).

	0	1	2	3	4	5	6	7	8	9	...
0	109.294356	97.709903	105.871828	102.38716	98.12077	135.066815	122.84465	133.423793	138.282193	136.024692	...

Εικόνα 41: Πρόβλεψη συνολικής κατανάλωσης κάθε πελάτη για τις επόμενες τρείς μέρες

Ομοίως, για να προβλέψουμε την κατανάλωση ανά δεκαπέντε λεπτά, ακολουθούμε την ίδια διαδικασία, χρησιμοποιώντας όμως το μοντέλο LR, καθώς αυτό ικανοποιούσε τα κριτήρια που αναφέρθηκαν παραπάνω με R-Squared τιμή 0.7751826785883065. Η πρόβλεψη της κατανάλωσης κάθε πελάτη, της επιλεγμένης συστάδας, ανά δεκαπέντε λεπτά, φαίνεται στην Εικόνα 42.

ment_135	165_segment_135	167_segment_135	168_segment_135	170_segment_135	171_segment_135	...	344_segment_135	352_segment_135	368_segment_135	Full Date	Time	Year	Day	Month	Day Name	Month Name
0.333143	0.266903	0.238536	0.313862	0.266275	0.216168	...	0.198084	0.175219	0.428623	2012-01-01	00:00:00	2012	1	1	Sunday	January
0.306738	0.323947	0.264743	0.319965	0.298865	0.218492	...	0.213306	0.195364	0.430826	2012-01-01	00:15:00	2012	1	1	Sunday	January
0.304836	0.254945	0.238848	0.331905	0.309886	0.259013	...	0.206109	0.250155	0.423415	2012-01-01	00:30:00	2012	1	1	Sunday	January
0.291334	0.245942	0.236168	0.296766	0.307489	0.260046	...	0.178803	0.262278	0.418109	2012-01-01	00:45:00	2012	1	1	Sunday	January
0.288750	0.216227	0.241689	0.289442	0.301802	0.275104	...	0.156064	0.270855	0.418731	2012-01-01	01:00:00	2012	1	1	Sunday	January

Εικόνα 42: Πρόβλεψη κατανάλωσης κάθε πελάτη (ανά 15') για τις επόμενες τρείς μέρες

3.1.2. Deep Neural Network - LSTM

Υλοποιήσαμε και μια LSTM προσέγγιση, ώστε να συγκρίνουμε τα αποτελέσματα των δύο μοντέλων. Αρχικά, επειδή τα δεδομένα μας είναι σε δισδιάστατη μορφή, κάνουμε έναν ανασχηματισμό ώστε να έρθουν σε τρισδιάστατη μορφή και να εισαχθούν στο μοντέλο. Έπειτα δημιουργούμε ένα απλό LSTM μοντέλο, το οποίο γίνεται και compile με optimizer τον Adam και με loss function το MSE (Εικόνα 43) . Στη συνέχεια εκπαιδεύουμε το μοντέλο με τα τρισδιάστατα σύνολα δεδομένων εκπαίδευσης, εκ των οποίων το 20% χρησιμοποιείται για validation (Εικόνα 44). Τέλος, αξίζει να σημειωθεί πως χρησιμοποιούμε πέντε epochs, καθώς η αριθμητική απώλεια (loss) είναι ικανοποιητική. Τέλος, κάνουμε κάποιες προβλέψεις μας πάνω στο, πλέον, τρισδιάστατο σύνολο ελέγχου X (Εικόνα 45), ώστε να δούμε πόσο απέχουν τα αποτελέσματα που θα δώσει το μοντέλο από τις πραγματικές τιμές.

```
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 lstm (LSTM)                 (None, 1, 288)            2102400

 dense (Dense)               (None, 1, 288)            83232

=================================================================
Total params: 2,185,632
Trainable params: 2,185,632
Non-trainable params: 0
_________________________________________________________________
```

Εικόνα 43: Μοντέλο LSTM

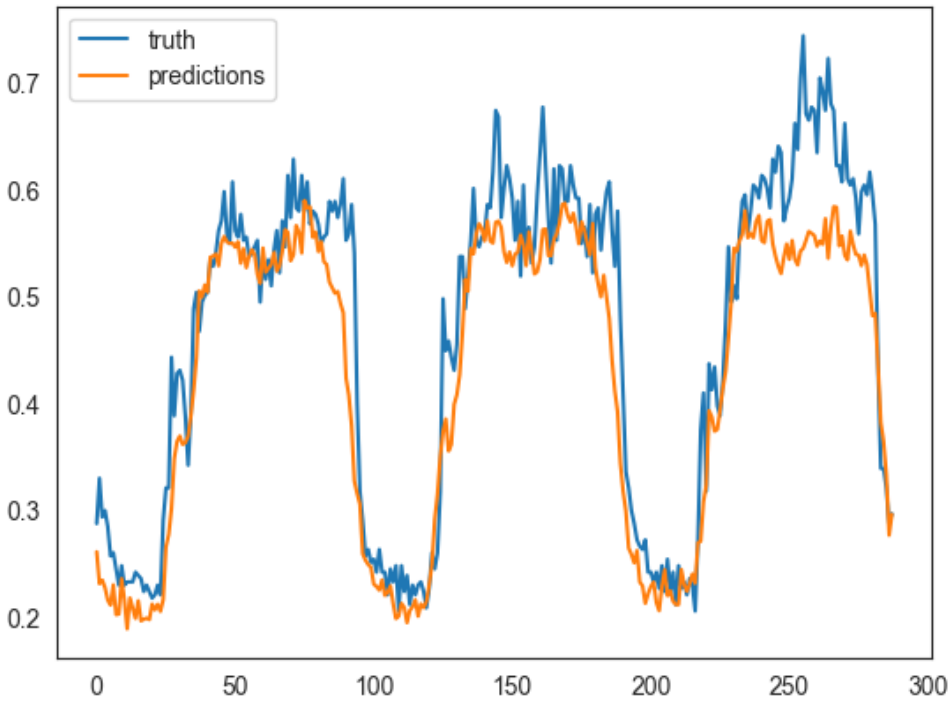
```
Epoch 1/5
255/255 [==============================] - 10s 30ms/step - loss: 0.0093
Epoch 2/5
255/255 [==============================] - 8s 30ms/step - loss: 0.0038
Epoch 3/5
255/255 [==============================] - 8s 31ms/step - loss: 0.0035
Epoch 4/5
255/255 [==============================] - 8s 30ms/step - loss: 0.0033
Epoch 5/5
255/255 [==============================] - 8s 30ms/step - loss: 0.0032
```

Εικόνα 44: Διαδικασία εκπαίδευσης μοντέλου LSTM

0	1	2	3	4	5	6	7	8	9	...	278	279	280	281	282	283	284	285	286	287
0.261507	0.231647	0.235015	0.227905	0.216274	0.211677	0.230559	0.202694	0.203244	0.236313	...	0.530055	0.506725	0.482283	0.484459	0.435737	0.382332	0.364885	0.333572	0.277023	0.296549
0.260562	0.240425	0.245343	0.243097	0.232723	0.219383	0.238491	0.212356	0.211779	0.247968	...	0.529136	0.518048	0.488913	0.492138	0.443014	0.388823	0.373586	0.337540	0.283184	0.308633
0.306567	0.284004	0.287318	0.280870	0.279022	0.251522	0.270780	0.250119	0.244823	0.286607	...	0.583030	0.568685	0.534240	0.528830	0.492338	0.431202	0.412163	0.375884	0.316371	0.341335
0.308224	0.284655	0.281369	0.278002	0.268985	0.255280	0.261464	0.249893	0.237315	0.270558	...	0.568099	0.556783	0.531389	0.547925	0.514615	0.463065	0.448807	0.408752	0.346263	0.361054
0.317413	0.289084	0.286882	0.275787	0.261749	0.241893	0.265245	0.246202	0.232592	0.262839	...	0.573206	0.566625	0.539675	0.569804	0.550129	0.518937	0.503326	0.456483	0.376874	0.376139
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
0.428356	0.420726	0.436125	0.453170	0.426138	0.421472	0.437063	0.428443	0.424834	0.455127	...	0.473984	0.443967	0.410144	0.457637	0.457926	0.446632	0.455471	0.450542	0.426076	0.455088
0.419468	0.394479	0.423692	0.447631	0.413050	0.409769	0.427366	0.412987	0.415875	0.447775	...	0.464003	0.431768	0.404382	0.445670	0.444586	0.442133	0.451309	0.445340	0.425499	0.457106
0.404359	0.383867	0.412034	0.435838	0.399847	0.398468	0.419993	0.398758	0.403579	0.436767	...	0.465998	0.444345	0.417948	0.453738	0.450610	0.437552	0.450765	0.433711	0.418705	0.445981
0.419130	0.397615	0.422341	0.442131	0.413831	0.410158	0.427530	0.407547	0.407076	0.441413	...	0.470990	0.448069	0.422004	0.454471	0.454984	0.443786	0.449353	0.437154	0.419745	0.450198
0.434246	0.413301	0.428427	0.451228	0.421575	0.423497	0.441234	0.418113	0.427608	0.457847	...	0.471333	0.451291	0.422021	0.453013	0.456308	0.446837	0.454289	0.441057	0.420697	0.461095

Εικόνα 45: Προβλέψεις LSTM πάνω στο σύνολο ελέγχου

Επιπλέον, αξίζει να σημειωθεί πως εφαρμόζουμε και κάποιες τεχνικές smoothing πάνω στα αποτελέσματα που έδωσε το μοντέλο LSTM, ώστε να εξαλείψουμε πιθανόν λανθασμένες τιμές οι οποίες δεν ανταποκρίνονται στις προβλέψεις. Ο τρόπος που υλοποιήθηκε η συγκεκριμένη διαδικασία είναι με μια απλή αντικατάσταση της λανθασμένης τιμής με την προηγούμενη τιμή. Για παράδειγμα οι παρακάτω τρείς τιμές (0,2 – 0,9 – 0,27) θα έχουν μορφή (0,2 – 0,2 – 0,27) μετά την εφαρμογή της τεχνικής smoothing. Με βάση λοιπόν τις παραπάνω προβλέψεις τα σκορ RMSE, R-Squared και MAPE, του μοντέλου, είναι 0.05148898828206212, 0.7202994844531351 και 0.17188475337817755 αντίστοιχα. Παρόλα αυτά, παρατηρούμε πως το MSE και R-Squared δεν είναι τόσο ικανοποιητικά σε σχέση με το MAPE σκορ, οπότε οπτικοποιούμε τις προβλέψεις για τον πρώτο πελάτη της επιλεγμένης συστάδας, για να υπάρξει μεγαλύτερη κατανόηση των αποτελεσμάτων (Εικόνα 46). Σύμφωνα με την παρακάτω εικόνα οι προβλέψεις, φαίνονται φυσιολογικές, γεγονός το οποίο καθιστά το μοντέλο έτοιμο να κάνει τις προβλέψεις για τις επόμενες τρεις μέρες.



Εικόνα 46: Προβλέψεις LSTM πάνω στο σύνολο ελέγχου. Όπως φαίνεται τα αποτελέσματα δεν απέχουν πολύ από τις πραγματικές τιμές.

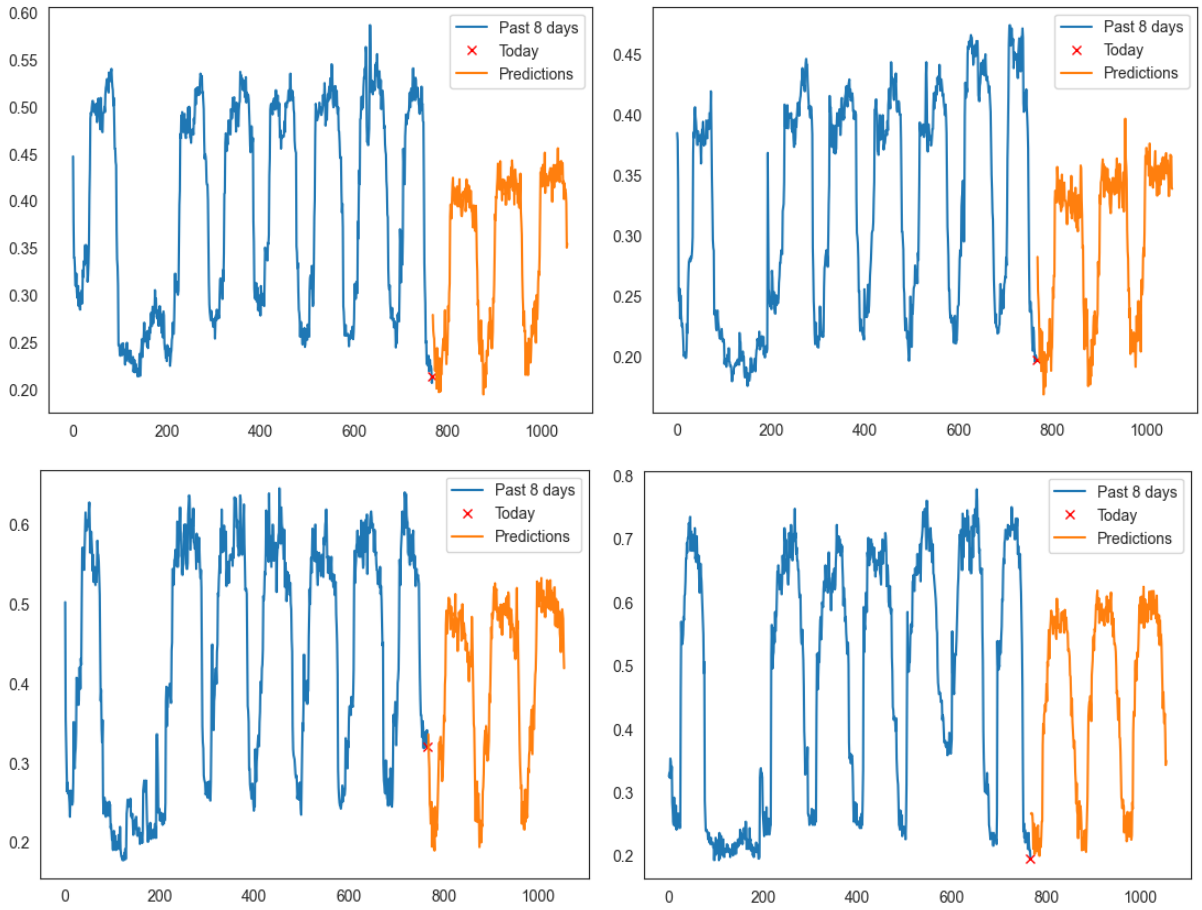
Έχει έρθει η ώρα να προβλέψουμε την συνολική κατανάλωση κάθε πελάτη για τις επόμενες τρείς μέρες με test set το X. Όπως αναφέρθηκε και προηγουμένως, φέρνουμε πρώτα το X σε τρισδιάστατη μορφή και το εισάγουμε στο μοντέλο ώστε να κάνει τις προβλέψεις. Αξίζει να σημειωθεί πως και σε αυτή τη διαδικασία προβλέψεων εφαρμόζονται τεχνικές smoothing, ώστε να έχουμε πιο ρεαλιστικά αποτελέσματα. Οι προβλέψεις ανά δεκαπέντε λεπτά και οι συνολικές προβλέψεις κάθε πελάτη, φαίνονται στην Εικόνα 47 και 48 αντίστοιχα. Τέλος, έχουμε οπτικοποιήσει τις προβλέψεις των επόμενων τριών ημέρων (ανά 15') για πέντε τυχαίους πελάτες (στην Εικόνα 49 φαίνονται οι τέσσερις από τους πέντε πελάτες), όπως επίσης την πρόβλεψη της συνολικής κατανάλωσης (Εικόνα 50).

	768	769	770	771	772	773	774	775	776	777	...	278	279	280	281	282	283	284	285	286	287
0	0.286957	0.236957	0.295652	0.445652	0.476087	0.489130	0.476087	0.463043	0.443478	0.428261	...	0.509408	0.507442	0.484251	0.474982	0.465838	0.438547	0.413765	0.385522	0.359179	0.333295
1	0.471591	0.409091	0.352273	0.323864	0.352273	0.289773	0.335227	0.272727	0.295455	0.289773	...	0.498654	0.480158	0.485683	0.483214	0.447606	0.436532	0.409093	0.369937	0.336292	0.323713
2	0.166667	0.145833	0.166667	0.159722	0.152778	0.159722	0.159722	0.145833	0.159722	0.152778	...	0.539553	0.506831	0.495748	0.478423	0.453099	0.418478	0.432089	0.373102	0.344957	0.341206
3	0.223881	0.223881	0.223881	0.283582	0.228856	0.228856	0.223881	0.223881	0.228856	0.223881	...	0.536883	0.514204	0.491213	0.467034	0.429818	0.419320	0.416731	0.381633	0.339052	0.345694
4	0.277664	0.240779	0.227459	0.234631	0.219262	0.215164	0.220287	0.205943	0.201844	0.204918	...	0.510532	0.499880	0.481087	0.449340	0.467358	0.436548	0.408747	0.365642	0.363942	0.324501
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
46	0.203205	0.183971	0.200434	0.192202	0.186715	0.181227	0.200434	0.186715	0.186715	0.192202	...	0.499307	0.496627	0.491917	0.468302	0.461705	0.419927	0.410644	0.370944	0.345031	0.347304
47	0.383292	0.405405	0.375921	0.385749	0.375921	0.407862	0.400491	0.393120	0.393120	0.388206	...	0.508582	0.484878	0.513459	0.472992	0.450935	0.421317	0.407479	0.390251	0.357206	0.343336
48	0.345216	0.319887	0.296435	0.259850	0.250469	0.214822	0.210131	0.210131	0.205441	0.205441	...	0.497107	0.483039	0.475538	0.460656	0.462864	0.437581	0.404722	0.378805	0.349787	0.328411
49	0.418336	0.411050	0.425622	0.404372	0.400729	0.282332	0.248937	0.237401	0.231937	0.229508	...	0.496621	0.494804	0.490522	0.478072	0.460612	0.442817	0.406882	0.380224	0.348371	0.330730
50	0.460956	0.468055	0.472788	0.459536	0.466162	0.494084	0.480360	0.476574	0.513488	0.524846	...	0.541646	0.523380	0.512820	0.488545	0.468888	0.445380	0.409598	0.393968	0.360182	0.336491

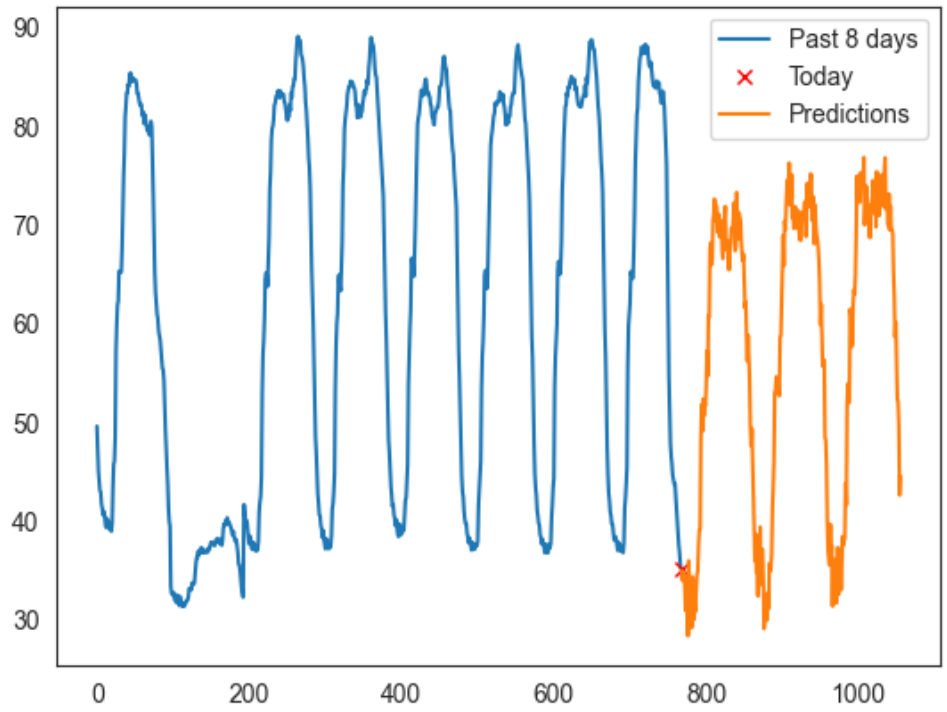
Εικόνα 47: Πρόβλεψη κατανάλωσης κάθε πελάτη (ανά 15') για τις επόμενες τρείς μέρες.

```
768    49.600720
769    47.044007
770    45.033697
771    43.923768
772    43.068269
         ...
283    64.820824
284    62.685089
285    57.716122
286    53.954765
287    50.373001
Length: 1056, dtype: float64
```

Εικόνα 48: Πρόβλεψη συνολικής κατανάλωσης κάθε πελάτη για τις επόμενες τρείς μέρες.



Εικόνα 49: Πρόβλεψη κατανάλωσης τεσσάρων τυχαίων πελατών (ανά 15') για τις επόμενες τρείς μέρες.



Εικόνα 50: Πρόβλεψη συνολικής κατανάλωσης κάθε πελάτη για τις επόμενες τρείς μέρες.

3.1.3. Σύγκριση Μοντέλων

Στον παρακάτω πίνακα φαίνεται η σύγκριση των επιδόσεων των μοντέλων LR και LSTM μέσω των τιμών RMSE και R-Squared.

	LR	LSTM
RMSE	0.04636102450284938	0.05148898828206212
R-Squared	0.7751826785883065	0.7202994844531351

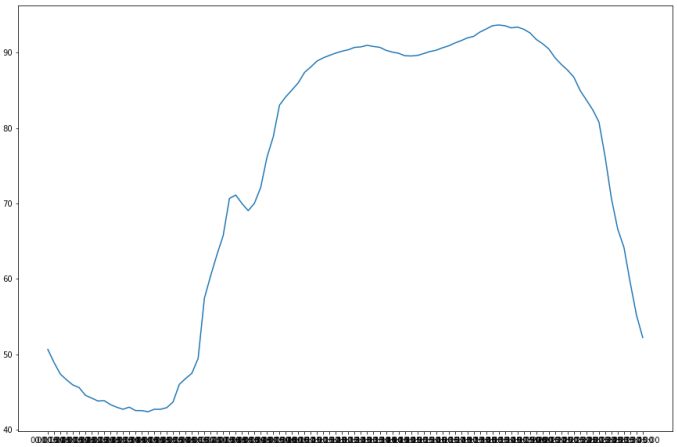
3.1.4. Απάντηση ερωτήματος d

Με βάση τις προβλέψεις για την κατανάλωση των πελατών τις επόμενες τρεις μέρες, παρατηρούμε μια συνολική μείωση, η οποία ωστόσο είναι σε φυσιολογικά πλαίσια και δεν καθιστά δεδομένη την μείωση κατανάλωσης κάθε μεμονωμένου πελάτη. Δηλαδή οι συνολικές τιμές κατανάλωσης κυμαίνονται στους ίδιους αριθμούς που έχουν καταγραφεί, απλά με μια μικρή μείωση. Για αυτό το λόγο, ένας ιδανικός τρόπος να ελαχιστοποιηθούν οι περίοδοι αιχμής (μέσα στην ημέρα) θα ήταν να γίνει ένας διαμοιρασμός της κατανάλωσης και σε βραδινές ώρες, για λόγους οικονομίας. Με αυτόν τον τρόπο το γράφημα της Εικόνας 50, δεν θα είχε τόσο έντονη κυματοειδή μορφή.

3.2. Βραχυπρόθεσμη Πρόβλεψη

3.2.1. Κλασσικοί Αλγόριθμοι Μηχανικής Μάθησης

Ομοίως με την μακροπρόθεσμη πρόβλεψη επιλέξαμε να δουλέψουμε με την τέταρτή συστάδα που δημιούργησε ο αλγόριθμος DBSCAN. Με βάση αυτή τη συστάδα θα προβλέψουμε πρώτον, την κατανάλωση κάθε πελάτη, ανά δεκαπέντε λεπτά, εντός τριών ωρών. Εφόσον γίνεται λόγος για ωριαία κατανάλωση, κρίναμε απαραίτητο να οπτικοποιούμε το άθροισμα της μέσης κατανάλωση των πελατών, ώστε να δούμε το εύρος των τιμών, σε περίπτωση που υπάρχουν φαινόμενα trend. Τα συμπεράσματα μας, βάσει της παρακάτω εικόνας είναι πως οι τιμές κατανάλωσης κυμαίνονται σε φυσιολογικά πλαίσια, με την παρουσία του φαινομένου trend αισθητή. Συγκεκριμένα υπάρχει απότομη αύξηση των τιμών μέσα στη μέρα και απότομη μείωση τους τις βραδινές ώρες. (Εικόνες 51, 52).




```
Time
00:00:00    50.676465
00:15:00    48.925919
00:30:00    47.404029
00:45:00    46.634611
01:00:00    45.961492
               ...
22:45:00    66.617255
23:00:00    64.185092
23:15:00    59.520868
23:30:00    55.215952
23:45:00    52.239631
Length: 96, dtype: float64
```

Εικόνες 51 & 52: 24ωρη αθροιστική χρονοσειρά (των ΜΟ) των πελατών της επιλεγμένης συστάδας

Καθώς οι χρονοσειρές του συνόλου δεδομένων είναι λίγες στο πλήθος, θα ακολουθήσουμε και πάλι την προσέγγιση επαύξησης των δεδομένων με την κατάτμησή τους μέσω ημι-επικαλυπτόμενων παραθύρων. Όπως και στην μακροπρόθεσμη πρόβλεψη, καλούμε τις συναρτήσεις divide_chunks() και get_labels() για κάθε καταναλωτή της επιλεγμένης συστάδας με μέγεθος segment 68,5 ώρες και μέγεθος label 3 ώρες, αντίστοιχα. Τα αποτελέσματα αυτής της διαδικασίας φαίνονται στην παρακάτω εικόνα (Εικόνα 53). Ωστόσο, παρατηρούμε, πάλι, πως το τελευταίο segment περιέχει κάποιες NaN τιμές, γεγονός, το οποίο είναι αναμενόμενο λόγω της διαδικασίας που ακολουθήσαμε, οπότε τις αφαιρούμε από το dataframe των αποτελεσμάτων (Εικόνα 54). Τέλος απομονώνουμε τις στήλες του τελευταίου segment της Εικόνας 53 σε ένα ξεχωριστό dataframe, το X (Εικόνα 55), καθώς το label του είναι μηδενικό και είναι αυτό που θέλουμε να προβλέψουμε.

90_segment_9	...	368_segment_758	368_segment_759	368_segment_760	368_segment_761	368_segment_762	368_segment_763	368_segment_764	368_segment_765	368_segment_766	368_segment_767
0.532609	...	0.550876	0.402272	0.571226	0.478940	0.426408	0.544250	0.431141	0.542830	0.421202	NaN
0.515217	...	0.511122	0.421675	0.570279	0.469948	0.435873	0.550402	0.432560	0.537624	0.420256	NaN
0.489130	...	0.485092	0.407004	0.543303	0.477520	0.458116	0.539991	0.430194	0.533838	0.432087	NaN
0.493478	...	0.451964	0.397066	0.542357	0.477520	0.459536	0.517274	0.439186	0.527212	0.432560	NaN
0.534783	...	0.453857	0.404165	0.563654	0.477993	0.456223	0.499290	0.451964	0.521060	0.417416	NaN

Εικόνα 53: Αποτελέσματα κατάτμησης χρονοσειρών μέσω επικαλυπτόμενων παραθύρων

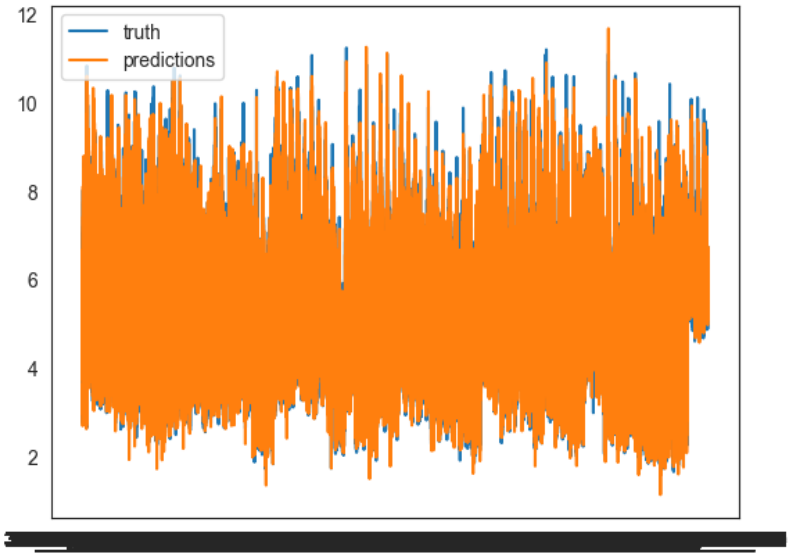
90_segment_9	...	368_segment_757	368_segment_758	368_segment_759	368_segment_760	368_segment_761	368_segment_762	368_segment_763	368_segment_764	368_segment_765	368_segment_766
6.315217	...	5.435400	5.761950	4.891150	6.745859	5.392806	6.088973	6.056318	5.150970	6.279224	0.000000
0.356522	...	0.470421	0.577378	0.456223	0.461903	0.426408	0.561761	0.453384	0.445812	0.508755	0.429248
0.332609	...	0.467108	0.572172	0.441552	0.470894	0.401325	0.570279	0.442972	0.477520	0.514434	0.418836
0.306522	...	0.467582	0.587317	0.464742	0.474681	0.398486	0.560814	0.459536	0.456223	0.486039	0.419782
0.319565	...	0.492664	0.581637	0.473734	0.489825	0.396593	0.566020	0.446758	0.465215	0.517274	0.424515

Εικόνα 54: Αποτελέσματα κατάτμησης χρονοσειρών μέσω επικαλυπτόμενων παραθύρων (χωρίς NaN τιμές)

	90_segment_766	125_segment_766	134_segment_766	136_segment_766	138_segment_766	165_segment_766	167_segment_766	168_segment_766	170_segment_766	171_segment_766	...	324_segment_766	325_segment_766	326_segment_766
0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.000000
1	0.245652	0.278409	0.131944	0.238806	0.195697	0.281335	0.232184	0.257420	0.300204	0.226284	...	0.245455	0.201922	0.195786
2	0.241304	0.289773	0.125000	0.228856	0.195697	0.254761	0.260673	0.250000	0.300204	0.226284	...	0.334545	0.212010	0.186757
3	0.245652	0.278409	0.131944	0.233831	0.193648	0.244152	0.240324	0.250000	0.300204	0.221474	...	0.300000	0.212010	0.198796
4	0.280435	0.272727	0.131944	0.228856	0.203893	0.254761	0.248464	0.250000	0.292839	0.207043	...	0.243636	0.186764	0.195786
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
270	0.315217	0.142045	0.173611	0.378109	0.372951	0.265422	0.301412	0.346679	0.292839	0.269627	...	0.221818	0.199400	0.183717
271	0.330435	0.153409	0.138889	0.278607	0.360656	0.244152	0.264743	0.272113	0.329444	0.240764	...	0.214545	0.227143	0.195786
272	0.328261	0.130682	0.131944	0.223881	0.363730	0.302551	0.276993	0.295034	0.322153	0.235905	...	0.212727	0.164065	0.156629
273	0.334783	0.125000	0.138889	0.218905	0.368852	0.265422	0.256603	0.295034	0.300204	0.235905	...	0.203636	0.171631	0.189767
274	0.319565	0.164773	0.131944	0.213930	0.361680	0.286639	0.244394	0.295034	0.262797	0.211853	...	0.190909	0.191808	0.180707

Εικόνα 55: Segment 766 (η πρώτη σειρά είναι τα labels που θέλουμε να προβλέψουμε)

Στη συνέχεια χωρίζουμε τα δεδομένα σε ετικέτες – δεδομένα εκπαίδευσης και ελέγχου, όπως και στην μακροπρόθεσμη πρόβλεψη και πλέον μπορούμε να δημιουργήσουμε και να εκπαιδεύσουμε τα μοντέλα μηχανικής μάθησης, ώστε να επιλέξουμε το καλύτερο και να το χρησιμοποιήσουμε για την πρόβλεψή μας. Συγκεκριμένα, για τους κλασσικούς αλγόριθμους μηχανικής μάθησης υλοποιήθηκε ένα μοντέλο γραμμικής παλινδρόμησης (Linear Regression – LR), όπου το εκπαιδεύουμε και τα αποτελέσματά του πάνω στο σύνολο ελέγχου φαίνονται στην Εικόνα 56. Όπως καταλαβαίνουμε, το μοντέλο LR έχει πολύ υψηλό ποσοστό επιτυχίας με R-Squared σκορ 0.9414237210670451 και RMSE σκορ 0.0951088824785518, δηλαδή οι προβλέψεις που έκανε είναι πολύ κοντά στις πραγματικές τιμές.



Εικόνα 56: Προβλέψεις LR πάνω στο σύνολο ελέγχου. Όπως φαίνεται τα αποτελέσματα δεν απέχουν πολύ από τις πραγματικές τιμές.

Τέλος, προβλέπουμε την κατανάλωση κάθε πελάτη (ανά 15') για τις επόμενες τρείς ώρες με test set το X, χρησιμοποιώντας το μοντέλο LR. Οι προβλέψεις ανά δεκαπέντε λεπτά και οι συνολικές προβλέψεις κάθε πελάτη, φαίνονται στην Εικόνα 57 και 58 αντίστοιχα. Τέλος, έχουμε οπτικοποιήσει τις προβλέψεις των επόμενων τριών ωρών (ανά 15') για πέντε τυχαίους πελάτες (στην Εικόνα 59 φαίνονται οι τέσσερις από τους πέντε πελάτες), όπως επίσης την πρόβλεψη της συνολικής κατανάλωσης (Εικόνα 60).

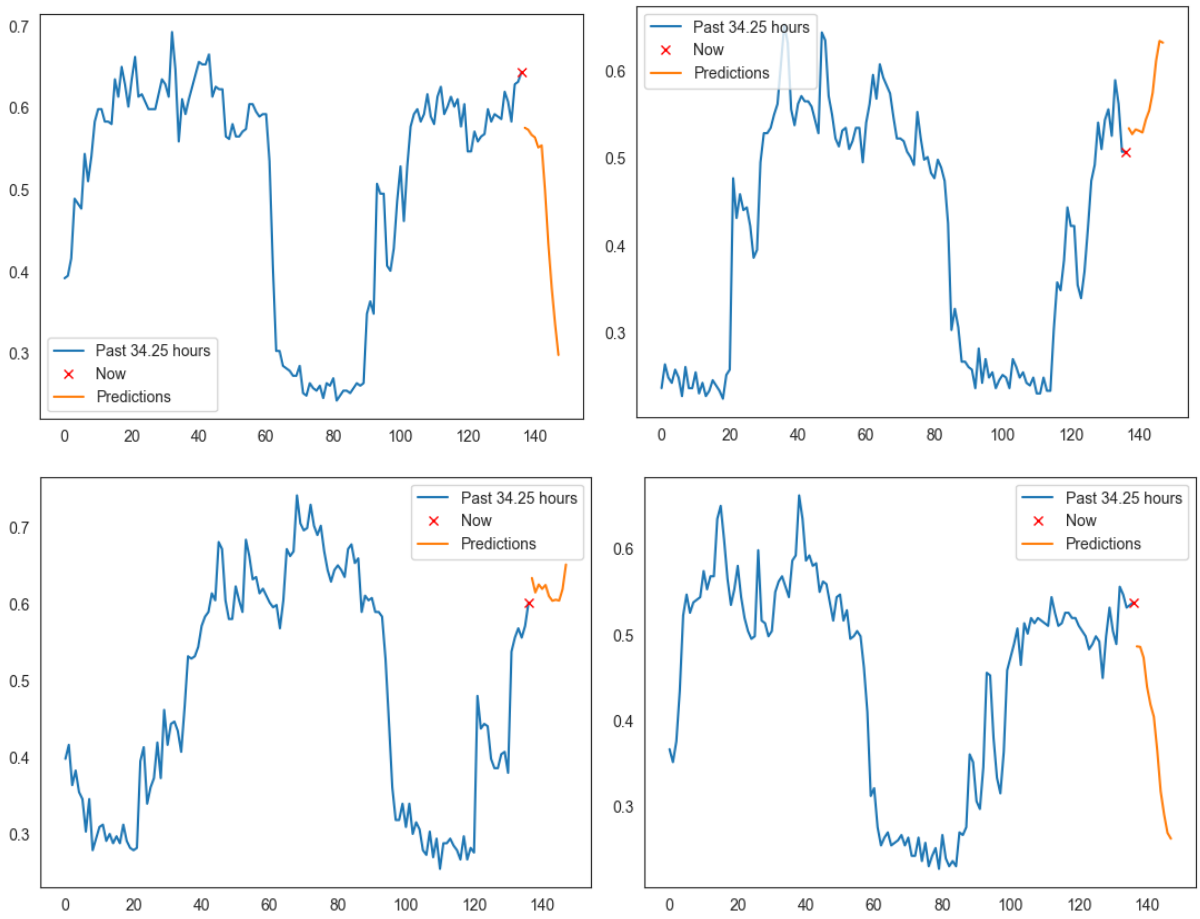
170_segment_766	171_segment_766	...	344_segment_766	352_segment_766	368_segment_766	Full Date	Time	Year	Day	Month	Day Name	Month Name
3.287032	2.705877	...	1.700329	2.103649	5.274780	2012-01-01	00:00:00	2012	1	1	Sunday	January
0.306531	0.232881	...	0.183666	0.288925	0.457780	2012-01-01	00:15:00	2012	1	1	Sunday	January
0.303600	0.237962	...	0.177803	0.287712	0.452963	2012-01-01	00:30:00	2012	1	1	Sunday	January
0.308321	0.233672	...	0.174795	0.286877	0.449627	2012-01-01	00:45:00	2012	1	1	Sunday	January
0.306305	0.222462	...	0.174788	0.262488	0.459273	2012-01-01	01:00:00	2012	1	1	Sunday	January

Εικόνα 57: Πρόβλεψη κατανάλωσης κάθε πελάτη (ανά 15') για τις επόμενες τρείς ώρες (LR)

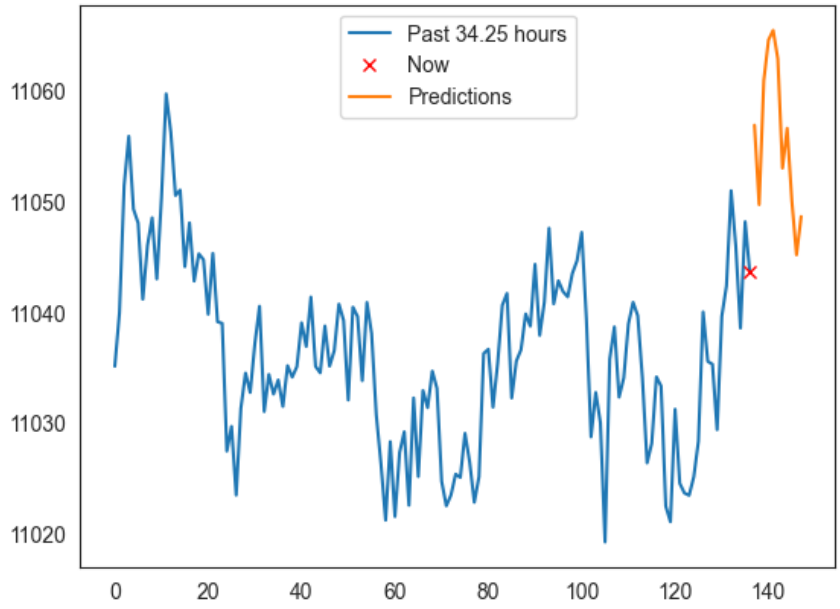
```
138    11035.149214
139    11039.964873
140    11051.442516
141    11055.884386
142    11049.318874
           ...
7      11053.007671
8      11056.606334
9      11050.088421
10     11045.196864
11     11048.654247
Length: 149, dtype: float64
```

Εικόνα 58: Πρόβλεψη συνολικής κατανάλωσης κάθε πελάτη για τις επόμενες τρείς ώρες (LR)

Οι προβλέψεις του μοντέλου γραμμικής παλινδρόμησης, όπως φαίνεται και παρακάτω, έχουν σε μεγάλο ποσοστό ακραίες μεταβατικές τιμές. Παρόλα αυτά, το trend που ακολουθούν (απότομη αύξηση το πρωί – απότομη μείωση το βράδυ) είναι σωστό και οι ακραίες τιμές οφείλονται στο γεγονός ότι προβλέπουμε τις επόμενες τρείς ώρες, οι οποίες αποτελούνται από μόνο δώδεκα σημεία δεδομένων.



Εικόνα 59: Πρόβλεψη κατανάλωσης τεσσάρων τυχαίων πελατών (ανά 15') για τις επόμενες τρείς ώρες (LR).



Εικόνα 60: Πρόβλεψη συνολικής κατανάλωσης κάθε πελάτη για τις επόμενες τρείς ώρες (LR).

3.2.2. Deep Neural Network - LSTM

Το δεύτερο μοντέλο που επιλέξαμε να υλοποιήσουμε, όπως και στην μακροπρόθεσμη πρόβλεψη είναι το LSTM Deep Neural Network, ώστε να συγκρίνουμε τα αποτελέσματά του με αυτά του μοντέλου της γραμμικής παλινδρόμησης. Ακολουθούμε την ίδια διαδικασία με την μακροχρόνια πρόβλεψη, δηλαδή κάνουμε ανασχηματισμό δεδομένων σε τρισδιάστατη μορφή, δημιουργούμε το LSTM μοντέλο (Εικόνα 61), το εκπαιδεύουμε με τα τρισδιάστατα σύνολα εκπαίδευσης (Εικόνα 62), χρησιμοποιώντας πέντε epochs και κάνουμε τις αντίστοιχες προβλέψεις (Εικόνα 62) ώστε να δούμε πόσο απέχουν τα αποτελέσματα που θα δώσει το μοντέλο από τις πραγματικές τιμές.

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 lstm (LSTM)                 (None, 1, 64)             83968

 dropout (Dropout)           (None, 1, 64)             0

 lstm_1 (LSTM)               (None, 1, 32)             12416

 dropout_1 (Dropout)         (None, 1, 32)             0

 dense (Dense)               (None, 1, 12)             396

=================================================================
Total params: 96,780
Trainable params: 96,780
Non-trainable params: 0
_________________________________________________________________
```

Εικόνα 61: Μοντέλο LSTM

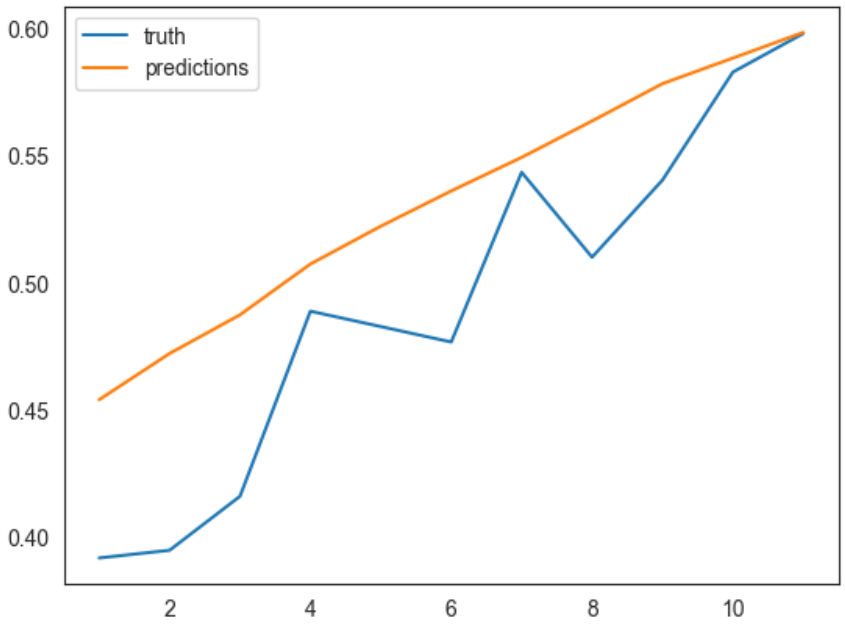
```
Epoch 1/5
1446/1446 [==============================] - 10s 4ms/step - loss: 0.1608
Epoch 2/5
1446/1446 [==============================] - 6s 4ms/step - loss: 0.0508
Epoch 3/5
1446/1446 [==============================] - 6s 4ms/step - loss: 0.0451
Epoch 4/5
1446/1446 [==============================] - 6s 4ms/step - loss: 0.0408
Epoch 5/5
1446/1446 [==============================] - 6s 4ms/step - loss: 0.0372
```

Εικόνα 62: Διαδικασία εκπαίδευσης μοντέλου LSTM

	0	1	2	3	4	5	6	7	8	9	10	11
0	6.843765	0.595784	0.599468	0.600758	0.601624	0.603762	0.600002	0.596227	0.594625	0.589004	0.585351	0.585457
1	2.998257	0.307568	0.317105	0.328841	0.339217	0.350941	0.360869	0.370265	0.380711	0.391189	0.402717	0.415558
2	6.746815	0.594504	0.598901	0.600992	0.602690	0.605610	0.602591	0.599787	0.598843	0.593843	0.590784	0.591465
3	4.673635	0.309819	0.304945	0.303192	0.298835	0.296992	0.292931	0.286054	0.285574	0.283931	0.284881	0.289440
4	6.288939	0.538609	0.541451	0.542936	0.543520	0.545627	0.542396	0.538848	0.538033	0.533780	0.531539	0.532820
...	...	...	...	...	...	...	...	...	...	...	...	...
23128	5.498737	0.451563	0.452686	0.454057	0.453965	0.455644	0.452979	0.449548	0.449660	0.447250	0.447118	0.450158
23129	5.744503	0.547766	0.556278	0.562766	0.569016	0.576312	0.578301	0.581008	0.584173	0.584368	0.585740	0.589934
23130	6.574056	0.508435	0.506055	0.503213	0.498982	0.496626	0.489318	0.480635	0.476790	0.469397	0.464812	0.464394
23131	5.362674	0.509032	0.517132	0.523852	0.530078	0.537524	0.540121	0.543181	0.546982	0.548285	0.550752	0.555896
23132	6.582034	0.536682	0.537135	0.536376	0.534668	0.534552	0.529266	0.523134	0.520718	0.514650	0.510951	0.511277

Εικόνα 63: Προβλέψεις LSTM πάνω στο σύνολο ελέγχου

Επιπλέον, αξίζει να σημειωθεί πως εφαρμόζουμε και στην βραχυπρόθεσμη πρόβλεψη εφαρμόζουμε τις ίδιες τεχνικές smoothing πάνω στα αποτελέσματα που έδωσε το μοντέλο LSTM, ώστε να εξαλείψουμε πιθανόν λανθασμένες τιμές οι οποίες δεν ανταποκρίνονται στις προβλέψεις. Το μοντέλο LSTM έχει σχετικά καλό ποσοστό επιτυχίας με R-Squared σκορ 0.8628120259005629, MSE σκορ 0.013464401608819338 και MAPE σκορ 0.15937961748146715. Αυτό φυσικά μπορεί να φανεί και στην Εικόνα 64.



Εικόνα 64: Προβλέψεις LSTM πάνω στο σύνολο ελέγχου. Όπως φαίνεται τα αποτελέσματα απέχουν ελάχιστα από τις πραγματικές τιμές.

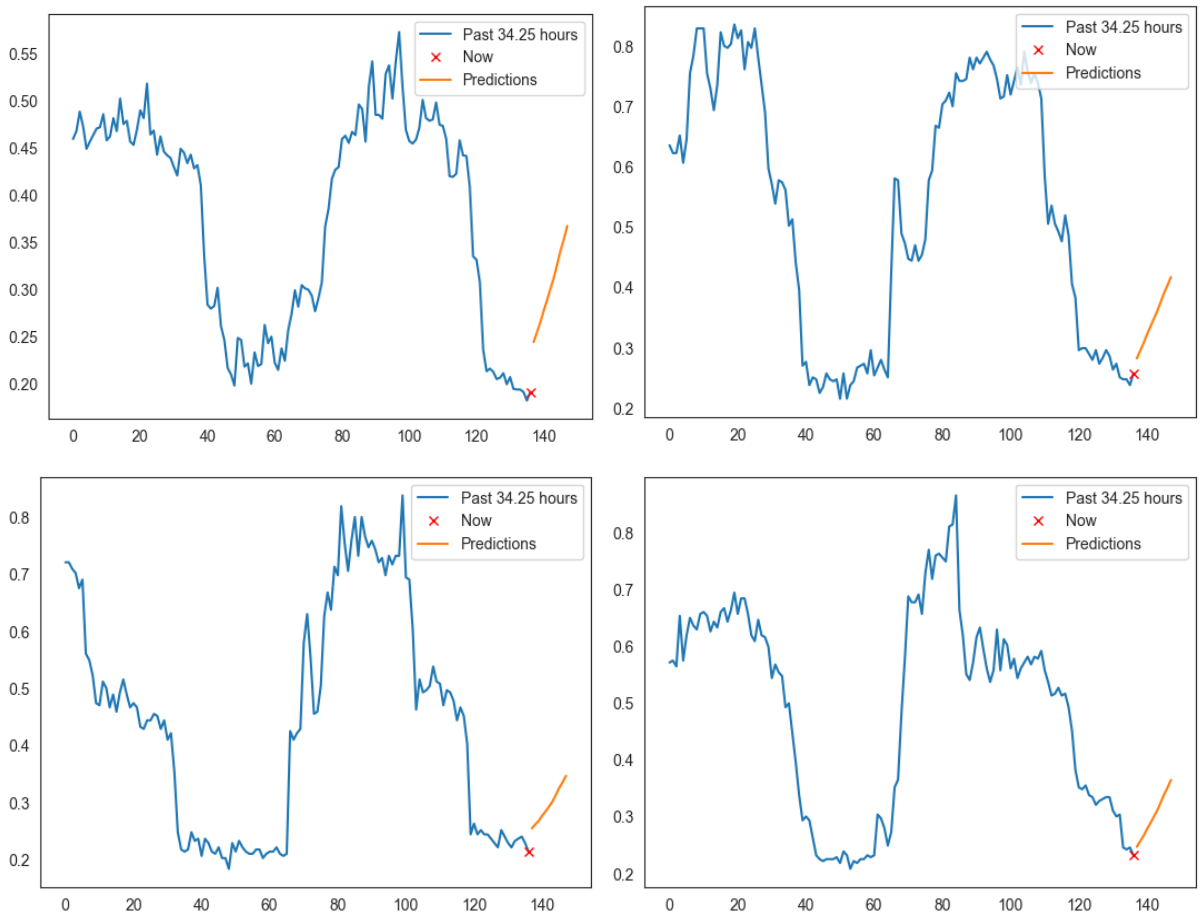
Έχει έρθει η ώρα να προβλέψουμε την συνολική κατανάλωση κάθε πελάτη για τις επόμενες τρείς ώρες με test set το X. Όπως αναφέρθηκε και προηγουμένως, φέρνουμε πρώτα το X σε τρισδιάστατη μορφή και το εισάγουμε στο μοντέλο ώστε να κάνει τις προβλέψεις. Αξίζει να σημειωθεί πως και σε αυτή τη διαδικασία προβλέψεων εφαρμόζονται τεχνικές smoothing, ώστε να έχουμε πιο ρεαλιστικά αποτελέσματα. Οι προβλέψεις ανά δεκαπέντε λεπτά και οι συνολικές προβλέψεις κάθε πελάτη, φαίνονται στην Εικόνα 65 και 66 αντίστοιχα. Τέλος, έχουμε οπτικοποιήσει τις προβλέψεις των επόμενων τριών ωρών (ανά 15') για πέντε τυχαίους πελάτες (στην Εικόνα 66 φαίνονται οι τέσσερις από τους πέντε πελάτες), όπως επίσης την πρόβλεψη της συνολικής κατανάλωσης (Εικόνα 67).

	0	1	2	3	4	5	6	7	8	9	...	253
0	0.319565	0.295652	0.273913	0.300000	0.373913	0.384783	0.382609	0.419565	0.406522	0.376087	...	0.356522
1	0.159091	0.187500	0.238636	0.198864	0.210227	0.170455	0.204545	0.232955	0.232955	0.448864	...	0.136364
2	0.145833	0.152778	0.152778	0.208333	0.590278	0.694444	0.652778	0.569444	0.562500	0.597222	...	0.375000
3	0.398010	0.293532	0.298507	0.393035	0.432836	0.447761	0.447761	0.437811	0.447761	0.467662	...	0.283582
4	0.221311	0.237705	0.245902	0.299180	0.327869	0.344262	0.337090	0.334016	0.331967	0.330943	...	0.366803
...	...	...	...	...	...	...	...	...	...	...	...	...
146	0.233386	0.255364	0.271854	0.269083	0.404544	0.451325	0.412803	0.334989	0.313038	0.291061	...	0.178483
147	0.277641	0.275184	0.250614	0.250614	0.375921	0.378378	0.373464	0.417690	0.378378	0.383292	...	0.171990
148	0.178236	0.188555	0.186679	0.193246	0.199812	0.203565	0.213884	0.217636	0.234522	0.235460	...	0.208255
149	0.230723	0.231937	0.248937	0.258045	0.190650	0.183971	0.180328	0.185792	0.197936	0.200364	...	0.211293
150	0.455277	0.436346	0.448651	0.450071	0.467582	0.480833	0.451964	0.484146	0.508755	0.514908	...	0.428301

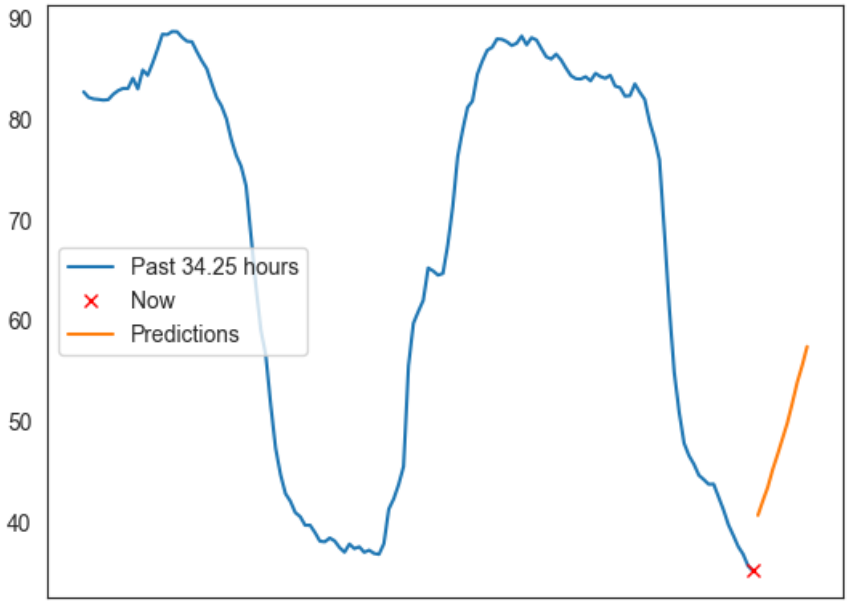
Εικόνα 65: Πρόβλεψη κατανάλωσης κάθε πελάτη (ανά 15') για τις επόμενες τρείς ώρες.

```
126    82.728102
127    82.182206
128    82.025243
129    81.977916
130    81.908911
          ...
7      48.445770
8      49.904991
9      51.396454
10     53.120033
11     55.043964
Length: 149, dtype: float64
```

Εικόνα 66: Πρόβλεψη συνολικής κατανάλωσης κάθε πελάτη για τις επόμενες τρείς ώρες.



Εικόνα 67: Πρόβλεψη κατανάλωσης τεσσάρων τυχαίων πελατών (ανά 15') για τις επόμενες τρείς ώρες.



Εικόνα 68: Πρόβλεψη συνολικής κατανάλωσης κάθε πελάτη για τις επόμενες τρείς ώρες.

3.2.3. Σύγκριση Μοντέλων

Στον παρακάτω πίνακα φαίνεται η σύγκριση των επιδόσεων των μοντέλων LR και LSTM μέσω των τιμών RMSE και R-Squared.

	LR	LSTM
RMSE	0.0951088247855177	0.10366115284418011
R-Squared	0.941423721067045	0.8801931033592564

3.2.4. Απάντηση ερωτήματος d

Με βάση τις προβλέψεις για την κατανάλωση των πελατών τις επόμενες τρεις ώρες, παρατηρούμε πως οι προβλέψεις κυμαίνονται, περίπου, στους ίδιους αριθμούς που έχουν καταγραφεί απλά με μια εμφανή, μικρή αύξηση. Παρόλα αυτά, δεν μπορούμε να κάνουμε κάποια πρόταση για την ελαχιστοποίηση περιόδων αιχμής καθώς οι προβλέψεις αφορούν ένα διάστημα τριών ωρών.