ΕΡΓΑΣΙΑ ΓΙΑ ΤΟ ΜΑΘΗΜΑ: "ΕΙΔΙΚΑ ΘΕΜΑΤΑ ΔΙΚΤΥΩΝ ΓΝΩΣΗΣ ΚΑΙ ΣΗΜΑΣΙΟΛΟΓΙΚΟΥ ΙΣΤΟΥ"

Στέλιος Μπατζιάκας

Εργασία για το μάθημα: «Ειδικά Θέματα Δικτύων Γνώσης και Σημασιολογικού Ιστού»

Table of Contents

Εισαγωγή 3
Μετατροπή των δεδομένων σε RDF 3
Περιγραφική Στατιστική 4
Περιγραφή των μεταβλητών 4
Γραφική Διερεύνηση 6
Έλεγχος κανονικότητας 9
Συσχέτιση Spearman 10
Μοντελοποιήση με παλινδρόμηση 12
Παλινδρόμηση για την πρόβλεψη της συνολικής παραγωγής σε σχέση με τις μέσες εργασίας 17
Ομαδοποίηση νομών σε σχέση με την παραγωγή κάθε προϊόντος 19
Ομαδοποίηση νομών σε σχέση με τις ημέρες εργασίας 23
Δεντρά regression για την μοντελοποίηση της συνολικής παραγωγής κάθε νομού 24
Classification της παραγωγής σε χαμηλή-ψηλή 26
Συμπεράσματα 29

Εισαγωγή

Τα δεδομένα που εξετάζονται στην παρούσα εργασία αφορούν τον αριθμό, την κυριότητα και την εργασία στις αγροτικές εκμεταλλεύσεις στους διάφορους νομούς της Ελλάδας για το έτος 2011. Προέρχονται από την Ελληνική Στατιστική υπηρεσία. Σύμφωνα με την ελληνική νομοθεσία ως αγροτική εκμετάλλευση ορίζεται "η μονάδα παραγωγής προς πώληση αγροτικών προϊόντων"[...]. Στις δραστηριότητες της γεωργικής εκμετάλλευσης περιλαμβάνεται παράλληλα με την παραγωγή των προϊόντων και η αποθήκευση, τυποποίηση, συσκευασία, εν γένει τοποθέτηση μέχρι του σταδίου της χονδρικής πώλησης, αποκλειστικά των προϊόντων που παράγει η ίδια γεωργική εκμετάλλευση, καθώς και η πρώτη χωρική μεταποίηση τους, ως και η διαχείριση ανανεώσιμων πηγών ενέργειας και η λειτουργία αγροτοτουριστικών μονάδων." (Βουλή των Ελληνών, 2010) (Νόμος 3874 (ΦΕΚ 151/Α/2010)).

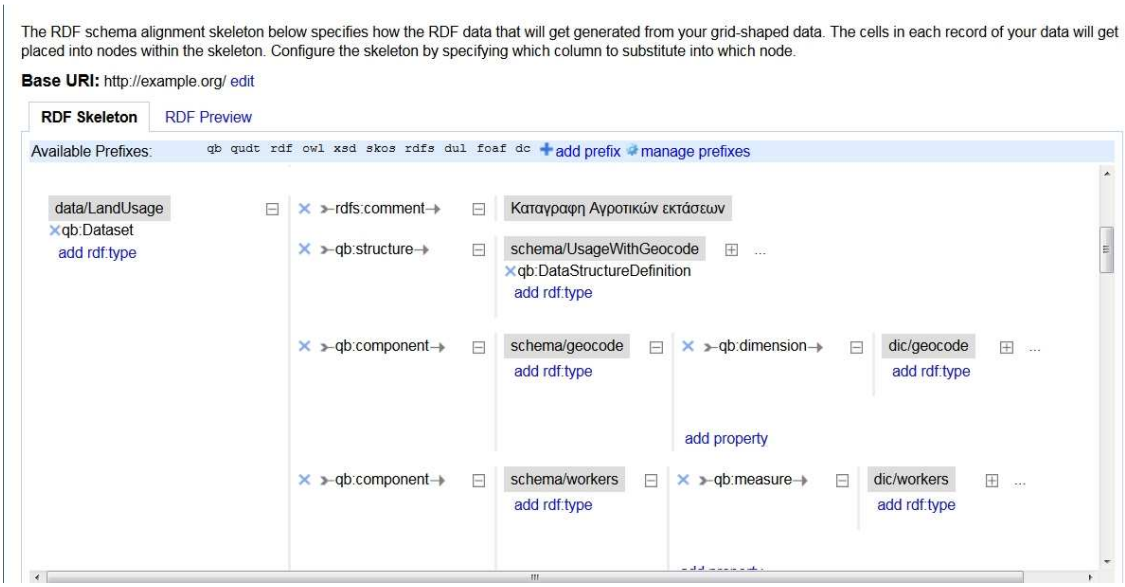
Η επεξεργασία των δεδομένων καθώς και η συγγραφή της παρουσίασης έγιναν αποκλειστικά στην R.

Μετατροπή των δεδομένων σε RDF

Το dataset μετασχηματίστηκε σε RDF format με το Open Refine, χρησιμοποιώντας κυρίως **Data Cube Vocabulary,** κάνοντας χρήση της ομώνυμης οντολογίας**.** Οι βασικές κλάσεις που χρειάστηκαν ήταν οι

- qb:structure για την δημιουργία του “σκελετού” του dataset
- qb:component για την περιγραφή της κάθε μεταβλητής
- qb:measure για τις τιμές των μεταβλητών
- qb:dimension για την περιγραφή της πολυπλοκότητας της κάθε μεταβλητής
- rdfs:comment για τα ονόματα των μεταβλητών

Τα linked data που προέκυψαν ανέβηκαν σε ένα local virtuoso repository και αντλήθηκαν στην R με την βιβλιοθήκη SPARQL. Ένα στιγμιότυπο από την διαδικασία RDFizing στο Open Refine φαίνεται στην παρακάτω εικόνα.



Περιγραφική Στατιστική

Στην παρούσα ενότητα θα παρουσιάσουμε με την βοήθεια διάφορων μέτρων περιγραφικής στατιστικής, το σύνολο των δεδομένων που αναλύθηκε.

Περιγραφή των μεταβλητών

Τα δεδομένα αποτελούνται από 23 μεταβλητές, από τις οποίες 14 είναι ποσοτικές και 2 ποιοτικές:

1. Ποιοτικές μεταβλητές:

- Το σύνολο των νομών στην Ελληνική επικράτεια
- Το σύνολο των περιφερειών της χώρας

Ποσοτικές μεταβλητές:

- Αριθμός αγροτικών εκμεταλλεύσεων στην Ελλάδα
    - Έκταση σε στρέμματα των αγροτικών εκτάσεων, των βοσκοτόπων και του κάθε νομού
    - Αριθμός εργαζομένων σε αγροτικές εκμεταλλεύσεις και σύνολο των ημερών που εργάστηκε.
    - Η αγροτική παραγωγή κάθε νομού σε προϊόντα που παράγονται από δέντρα

Διακρίνονται 4 κατηγορίες εργαζομένων: Οι ιδιοκτήτες και οι οικογένειες τους, οι μόνιμοι και οι εποχιακοί εργαζόμενοι καθώς και μια κατηγορία που περιέχει οποιαδήποτε άλλη περίπτωση όπως πχ. την εθελοντική εργασία στα πλαίσια αγροτουρισμού και την δωρεάν αλληλοβοήθεια μεταξύ αγροτών. Στην συνέχεια της εργασίας , ιδιαίτερα σε πίνακες θα αναφέρονται ως εργαζόμενοι κατηγορίας 1,2,3 και 4 αντίστοιχα για λόγους συντομίας.

Τα αγροτικά προϊόντα προέρχονται από διάφορα δέντρα τα οποία μπορούμε να τα χωρίσουμε σε 7 κατηγορίες: Μηλοειδή, Ροδακινοειδή, Ακρόδρυα, Εσπεριδοειδή, Ελιές, Αμπέλια και Άλλα. Στην συνέχεια της εργασίας , ιδιαίτερα σε πίνακες θα αναφέρονται ως Μήλα, Ροδάκινα, Καρποί με κέλυφος, Πορτοκάλια, Ελιές και Σταφύλια για λόγους συντομίας.

Περιγραφικά στατιστικά στοιχεία σχετικά με τις εκμεταλλεύσεις :

	mean	sd	median	min	max	range	skew	kurtosis	Se
Εκμεταλλεύσεις	14176.6	8456	13207	2266	42674	40408	1.045	1.083	1184.0
Με εκτάσεις	14055.4	8428	13036	2262	42647	40385	1.060	1.138	1180.2
Μέγεθος εκτάσεων	682.0	431	559	40	1818	1778	0.657	-0.385	60.3
Βοσκοτόπια	90.4	166	22	0	689	689	2.529	5.462	23.2
Πληθυσμός	214932.4	527591	112615	22506	3761810	3739304	6.026	37.183	73877.5
Έκταση	2581.9	1240	2519	356	5461	5105	0.363	-0.405	173.6

Περιγραφικά στατιστικά στοιχεία σχετικά με τον αριθμό εργαζομένων και τις ημέρες εργασίας :

	mean	sd	median	min	max	range	skew	kurtosis	se
Ιδιοκτήτες	23888	15903	21263	3451	77182	73731	1.22	1.599	2226.9
Μέρες	2287874	1624718	1900449	216840	7765688	7548848	1.13	1.236	227506
Σταθεροί εργαζόμενοι	491	412	385	5	1723	1718	1.19	0.971	57.7
Μέρες	101228	99825	77888	925	515292	514367	2.03	4.824	13978.3
Εποχιακοί	17561	17177	12369	313	69938	69625	1.25	0.714	2405.2
Μέρες	244781	245265	166752	2463	975190	972727	1.34	1.147	34344.0
Άλλοι	15094	14035	9555	538	58842	58304	1.35	1.556	1965.3
Μέρες	49380	40951	36333	2636	172015	169379	1.29	1.377	5734.3

Περιγραφικά στατιστικά στοιχεία σχετικά με την παραγωγή διάφορων αγροτικών προϊόντων:

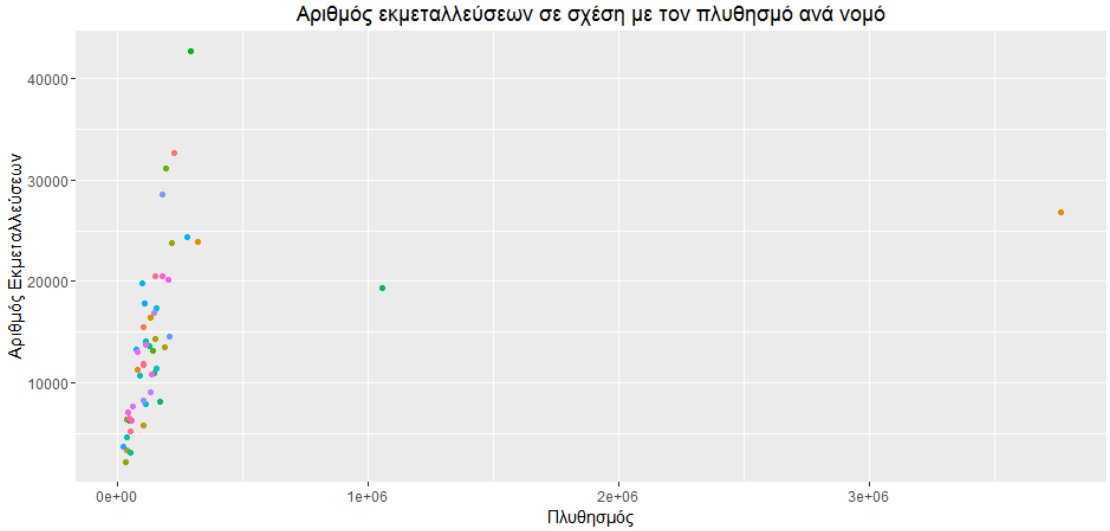
	mean	sd	median	min	max	range	skew	kurtosis	se
Μηλοειδή	6328	17045	461	0	86279	86279	3.44	11.62	2387
Ροδακινοειδή	16323	66018	550	0	343915	343915	4.52	19.11	9244
Ακρόδρυα	1612	2826	531	0	16829	16829	3.43	14.30	396
Εσπεριδοειδή	21602	59138	312	0	305034	305034	3.43	11.62	8281
Αλλά	3886	9270	291	0	51312	51312	3.37	12.29	1298
Σταφύλια	20552	38823	5637	0	210752	210752	3.04	10.23	5436
Ελιές	36742	58141	8781	0	268741	268741	2.26	5.03	8141

Γραφική Διερεύνηση

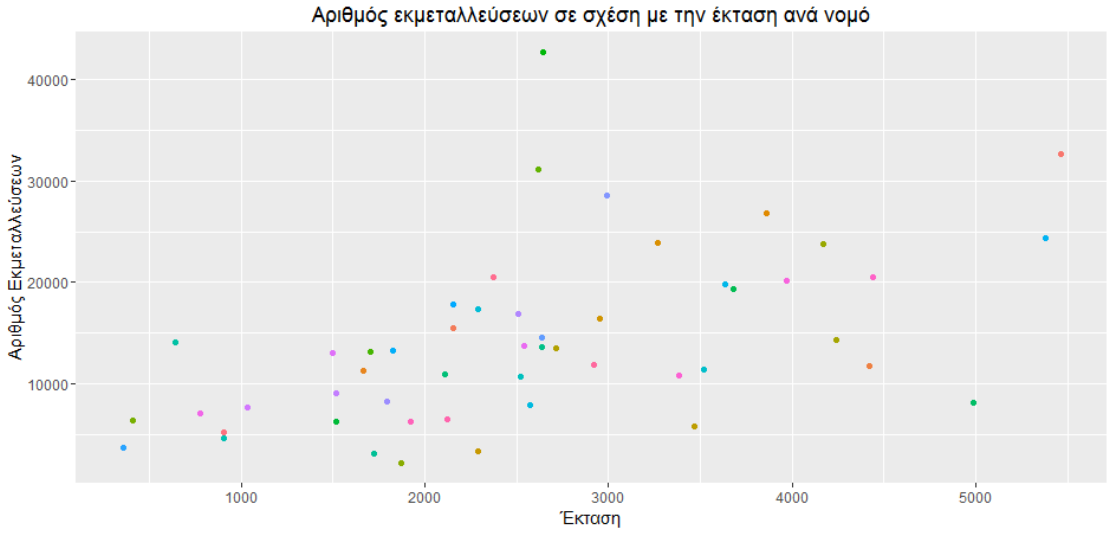
Στην συνέχεια δημιουργήσαμε μια σειρά από γραφήματα ώστε να κατανοήσουμε καλύτερα την συμπεριφορά των δεδομένων. Για την καλύτερη θέαση των γραφημάτων προτείνεται η χρήση του application που δημιουργήθηκε για την εργασία.

Στο πρώτο γράφημα παρατηρούμε ότι γενικά όσο αυξάνεται ο πληθυσμός αυξάνεται και ο αριθμός των εκμεταλλεύσεων με εξαίρεση τα δύο μεγαλύτερα

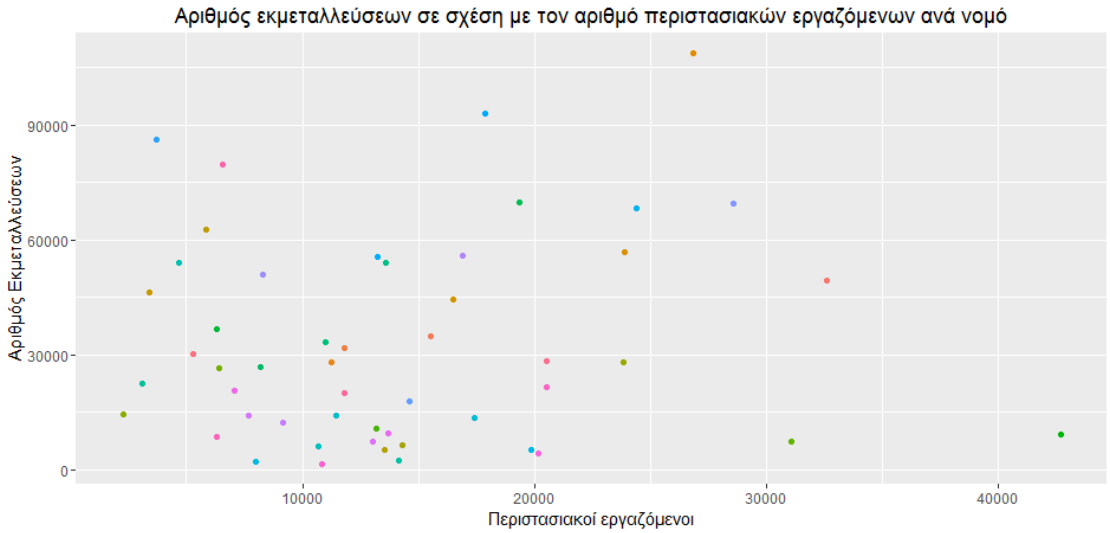
αστικά κέντρα την Αθήνα και την Θεσσαλονίκη.



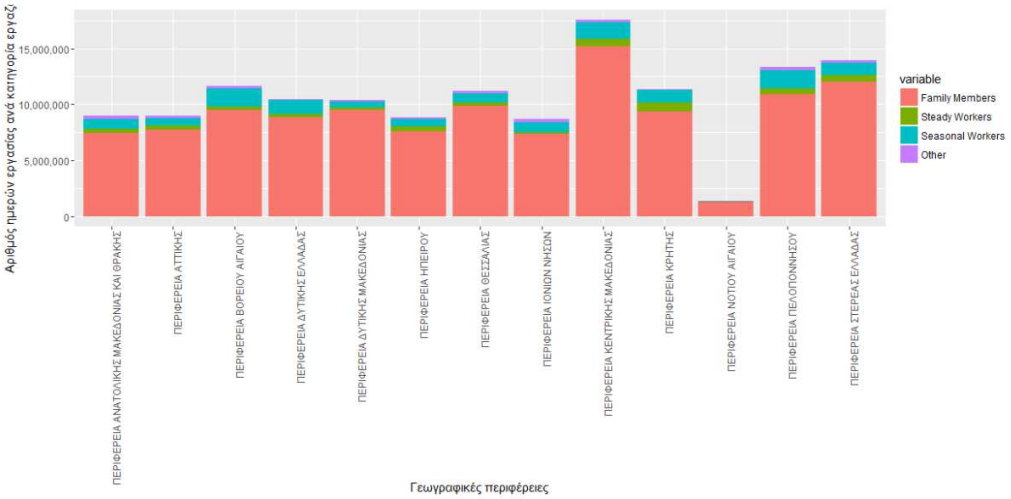
Από το δεύτερο γράφημα φαίνεται ο αριθμός των εκμεταλλεύσεων να μην έχει σχέση με το πόσο μεγάλη έκταση έχουν αυτές.



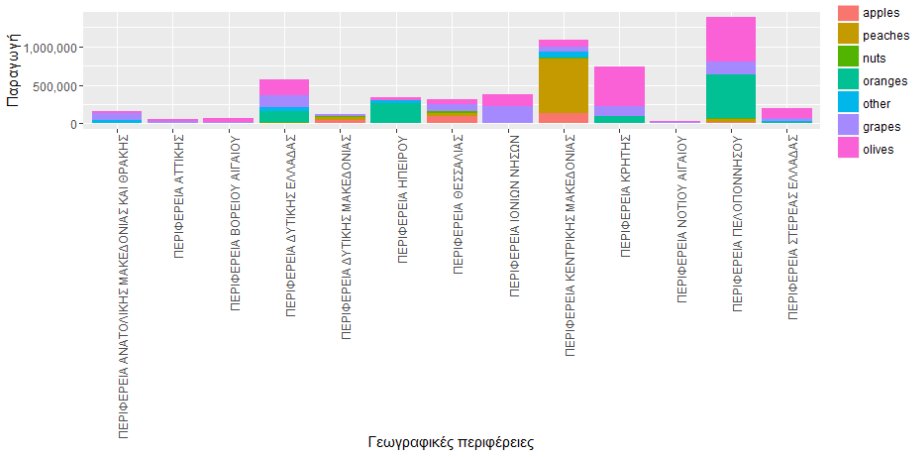
Από το τρίτο γράφημα φαίνεται ο αριθμός των εκμεταλλεύσεων να μην έχει σχέση με το πόσοι εποχιακοί εργαζόμενοι χρειάστηκαν για να καλυφθούν οι ανάγκες της παραγωγής.



Από τα δύο επόμενα γραφήματα διαπιστώνουμε πως η μεγαλύτερη πλειοψηφία των εργαζομένων στον αγροτικό τομέα έχει οικογενειακούς δεσμούς με τον ιδιοκτήτη της εκάστοτε εκμετάλλευσης και οι περαιτέρω ανάγκες καλύπτονται ως επί το πλείστον από εποχιακούς εργάτες ενώ οι σταθεροί εργαζόμενοι είναι ελάχιστοι κατά αναλογία. Το ίδιο φαίνεται και στις ημέρες εργασίας.



Όσον αφορά την παραγωγή, το μεγαλύτερο κομμάτι γίνεται κατά φθίνουσα σειρά στην Πελοπόννησο, στην Κεντρική Μακεδονία, στην Κρήτη και στην Δυτική Μακεδονία με την κάθε περιοχή να έχει σημαντική «υπεροχή» στην παραγωγή ενός προϊόντος.



Έλεγχος κανονικότητας

Με την βοήθεια του test Shapiro-Wilk, διαπιστώνουμε πως οι περισσότερες από τις μεταβλητές μας δεν ακολουθούν κανονική κατανομή. Η μοναδική για τις οποία η τιμή του p value είναι μεγαλύτερη από .05 οπότε και μπορούμε να υποθέσουμε κανονικότητα είναι η έκταση του κάθε νομού με statistic = 0.976 (p = 0.382). Για τον λόγο αυτό δεν θα χρησιμοποιήσουμε στατιστικές διαδικασίες που προϋποθέτουν κανονική κατανομή.

Συσχέτιση Spearman

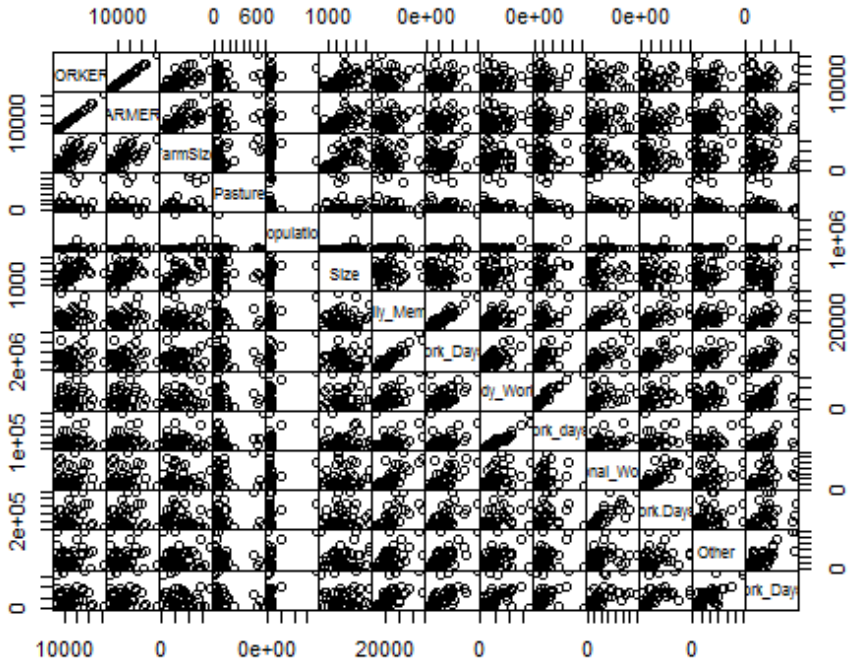
Στον παρακάτω πίνακα, παρουσιάζονται οι στατιστικά σημαντικές ($p < 0.05$) συσχετίσεις Spearman για τις διάφορες μεταβλητές. Κρατήσαμε τις τιμές για πάνω από 0,5 κατά απόλυτο τιμή . Παρατηρούμε διάφορες τετριμμένες συσχετίσεις όπως πχ. ανάμεσα στην έκταση του Νομού με την έκταση των εκμεταλλεύσεων αλλά και κάποιες ενδιαφέρουσες όπως ανάμεσα στους αριθμούς των διάφορων παραγόμενων ειδών που μπορεί να οφείλονται στην επιλογή από καλλιέργειες που αποδίδουν σοδειά σε διαφορετικές εποχές και κατά συνέπεια εισόδημα όλο τον χρόνο στους αγρότες πχ. τα μήλα (Φθινόπωρο) και τα ροδάκινα (Καλοκαίρι).

Μια ακόμα ενδιαφέρουσα υπόθεση προκύπτει από τις συσχετίσεις ανάμεσα στις ημέρες εργασίας των εργαζομένων με οικογενειακή σχέση και όλων των υπολοίπων κατηγοριών. Η θετική συσχέτιση ενδεχομένως να φανερώνει τον τρόπο λειτουργίας των γεωργικών εκμεταλλεύσεων όπου μόνο αν δεν είναι αρκετά τα εργατικά χέρια από το οικογενειακό περιβάλλον προσλαμβάνονται άλλοι εργαζόμενοι είτε εποχιακοί είτε μόνιμοι.

Μεταβλητή 1	Μεταβλητή 2	value
Εκμεταλλεύσεις	Εκμεταλλεύσεις με έκταση	1.000
Εκμεταλλεύσεις	Έκταση εκμεταλλεύσεων	0.736
Εκμεταλλεύσεις	Πληθυσμός	0.831
Εκμεταλλεύσεις	Έκταση Νομού	0.596
Εκμεταλλεύσεις	Σταφύλια	0.508
Εκμεταλλεύσεις	Ελιές	0.664
Εκμεταλλεύσεις με έκταση	Έκταση εκμεταλλεύσεων	0.733
Εκμεταλλεύσεις με έκταση	Πληθυσμός	0.827
Εκμεταλλεύσεις με έκταση	Έκταση Νομού	0.591
Εκμεταλλεύσεις με έκταση	Σταφύλια	0.502
Εκμεταλλεύσεις με έκταση	Ελιές	0.666
Έκταση εκμεταλλεύσεων	Πληθυσμός	0.651
Βοσκοτόπια	Πορτοκάλια	0.521
Πληθυσμός	Έκταση Νομού	0.725
Έκταση Νομού	Έκταση εκμεταλλεύσεων	0.684
Εργαζόμενοι κατ. 1	Μέρες εργασίας κατ. 1	0.921
Εργαζόμενοι κατ. 1	Εργαζόμενοι κατ. 2	0.566
Εργαζόμενοι κατ. 1	Μέρες εργασίας κατ. 2	0.551

Εργαζόμενοι κατ. 1	Εργαζόμενοι κατ. 3	0.859
Εργαζόμενοι κατ. 1	Μέρες εργασίας κατ. 3	0.870
Εργαζόμενοι κατ. 1	Εργαζόμενοι κατ. 4	0.738
Εργαζόμενοι κατ. 1	Μέρες εργασίας κατ. 4	0.839
Μέρες εργασίας κατ. 1	Εργαζόμενοι κατ. 2	0.577
Μέρες εργασίας κατ. 1	Μέρες εργασίας κατ. 2	0.590
Μέρες εργασίας κατ. 1	Εργαζόμενοι κατ. 3	0.764
Μέρες εργασίας κατ. 1	Μέρες εργασίας κατ. 3	0.787
Μέρες εργασίας κατ. 1	Εργαζόμενοι κατ. 4	0.741
Μέρες εργασίας κατ. 1	Μέρες εργασίας κατ. 4	0.796
Εργαζόμενοι κατ. 2	Μέρες εργασίας κατ. 2	0.972
Εργαζόμενοι κατ. 2	Εργαζόμενοι κατ. 3	0.588
Εργαζόμενοι κατ. 2	Μέρες εργασίας κατ. 3	0.627
Εργαζόμενοι κατ. 2	Εργαζόμενοι κατ. 4	0.524
Εργαζόμενοι κατ. 2	Μέρες εργασίας κατ. 4	0.509
Μέρες εργασίας κατ. 2	Εργαζόμενοι κατ. 3	0.560
Μέρες εργασίας κατ. 2	Μέρες εργασίας κατ. 3	0.611
Μέρες εργασίας κατ. 2	Εργαζόμενοι κατ. 4	0.526
Εργαζόμενοι κατ. 3	Μέρες εργασίας κατ. 3	0.941
Εργαζόμενοι κατ. 3	Εργαζόμενοι κατ. 4	0.635
Εργαζόμενοι κατ. 3	Μέρες εργασίας κατ. 4	0.754
Μέρες εργασίας κατ. 3	Εργαζόμενοι κατ. 4	0.617
Μέρες εργασίας κατ. 3	Μέρες εργασίας κατ. 4	0.790
Εργαζόμενοι κατ. 4	Μέρες εργασίας κατ. 3	0.617
Μήλα	Ροδάκινα	0.796
Μήλα	Καρποί με κέλυφος	0.774
Ροδάκινα	Καρποί με κέλυφος	0.628
Πορτοκάλια	Ελιές	0.703

Τέλος στον παρακάτω πίνακα μπορούμε να δρούμε και γραφικά τις σχέσεις ανάμεσα στις διάφορες μεταβλητές.



Μοντελοποιήση με παλινδρόμηση

Θα δοκιμάσουμε να μοντελοποιήσουμε τον αριθμό των αγροτικών εκμεταλλεύσεων σε συνάρτηση με τον αριθμό ωρών εργασίας από κάθε ομάδα. Επειδή από την φύση της η μεταβλητή που περιγράφει τις αγροτικές εκμεταλλεύσεις είναι ποσοτική διακριτή (count data), επιλέχθηκε η μοντελοποίηση με Poisson regression.

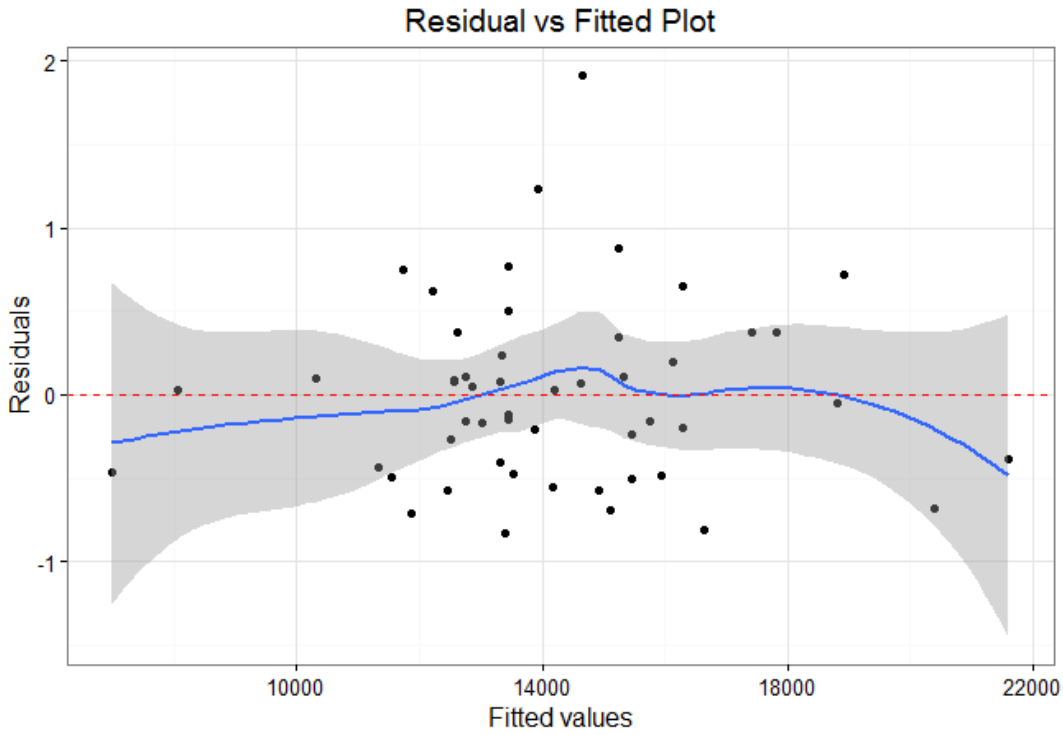
Δυστυχώς δεν έχουμε καλά αποτελέσματα.

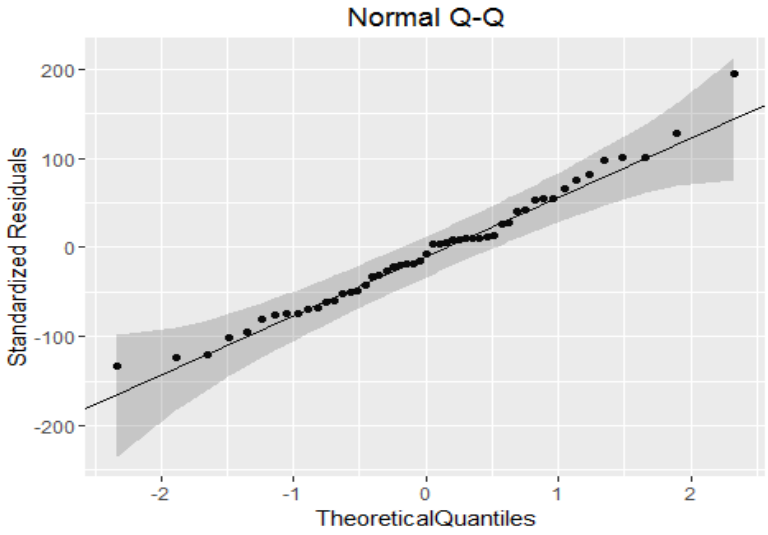
```
## glm(formula = ΕΚΜΕΤΑΛΛΕΎΣΕΙΣ ~ Μέρες εργασίας κατ. 1 + Μέρες εργασία
ς      κατ.      2      +      Μέρες      εργασίας      κατ.      3      +
##         Μέρες  εργασίας  κατ.  4,  family  =  "poisson",  data  =  df)
##
##                            Deviance                                Residuals:
##          Min               1Q      Median              3Q             Max
##   -129.23         -54.48           -6.78           33.17         187.63
##
##                                                                Coefficients:
##                              Estimate  Std.  Error  z  value  Pr(>|z|)
##  (Intercept)     9.45e+00        2.16e-03     4367.4        <2e-16  ***
## Μέρες εργασίας κατ. 1    1.22e-07    1.20e-09    101.8    <2e-16 ***
## Μέρες εργασίας κατ. 2    3.28e-07    1.28e-08     25.7    <2e-16 ***
## Μέρες εργασίας κατ. 3    2.92e-07    6.44e-09     45.2    <2e-16 ***
```

```
## Μέρες εργασίας κατ. 4   -5.88e-06    4.73e-08   -124.3    <2e-16 ***
##                                                                  ---
## Signif. codes:   0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##   (Dispersion parameter for poisson family taken to be 1)
##
##           Null deviance: 239137    on 50    degrees of freedom
##   Residual deviance: 213822     on 46     degrees of freedom
##                              AIC:                              214403
##
## Number of Fisher Scoring iterations: 5
```

Παρόλο που παρατηρούμε στατιστικά σημαντικές τιμές, η τιμή της εκτίμησης για το intercept είναι πάρα πολύ μεγάλη σε σχέση με τις εκτιμήσεις για τις μέρες εργασίας.

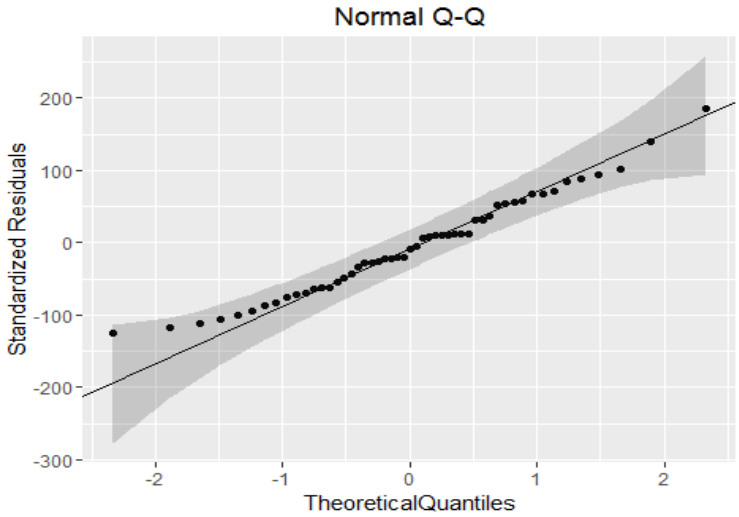
Τα υπόλοιπα της παλινδρόμησης ναι μεν παρουσιάζουν κανονικότητα, αλλά όχι καλό fit:

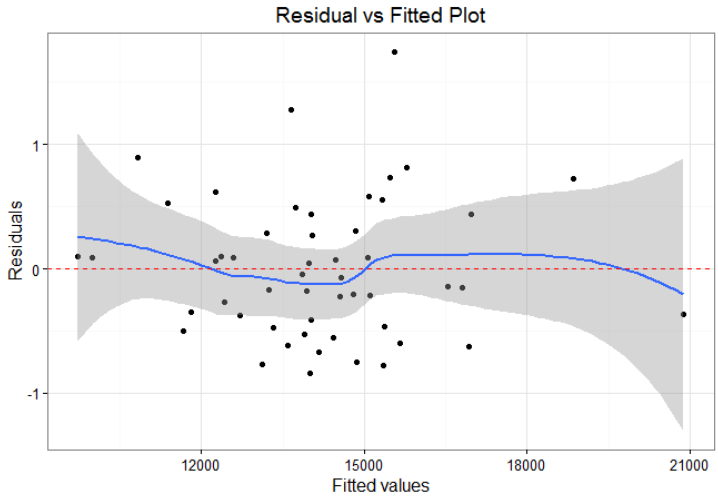




Θα δοκιμάσουμε την αντίστοιχη διαδικασία, αλλά αυτή την φορά με την μέση τιμή εργασίας ανά εργαζόμενο, για κάθε κατηγορία, για κάθε νομό, παρατηρούμε όμως την ίδια συμπεριφορά:

```
##
##                                                    Call:
## glm(formula = ΕΚΜΕΤΑΛΛΕΥΣΕΙΣ ~ Μέρες εργασίας κατ. 1 + Μέρες εργασία
ς      κατ.      2      +      Μέρες      εργασίας      κατ.      3      +
##         Μέρες  εργασίας  κατ.  4,  family  =  "poisson",  data  =  df2)
##
##                            Deviance                            Residuals:
##          Min              1Q      Median              3Q             Max
##   -123.35         -58.36           -8.79           42.10         178.56
##
##                                                  Coefficients:
##                              Estimate  Std.  Error  z  value  Pr(>|z|)
##  (Intercept)     9.58e+00        7.33e-03   1306.04        <2e-16  ***
## Μέρες εργασίας κατ. 1    4.46e-03    4.26e-05   104.52    <2e-16 ***
## Μέρες εργασίας κατ. 2   -1.64e-03    3.12e-05   -52.75    <2e-16 ***
## Μέρες εργασίας κατ. 3   -1.93e-03    2.06e-04    -9.37    <2e-16 ***
## Μέρες εργασίας κατ. 4   -2.08e-02    4.17e-04   -49.90    <2e-16 ***
##                                                                  ---
## Signif. codes:   0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##  (Dispersion   parameter   for   poisson   family   taken   to   be   1)
##
##           Null  deviance:  239137    on  50    degrees  of  freedom
##  Residual   deviance:   224955     on   46     degrees   of   freedom
##                              AIC:                                225537
##
## Number of Fisher Scoring iterations: 5
```

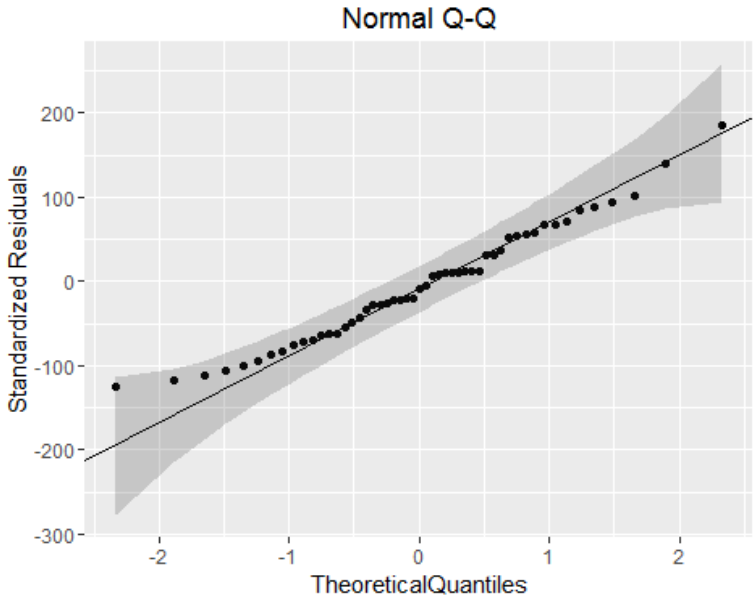


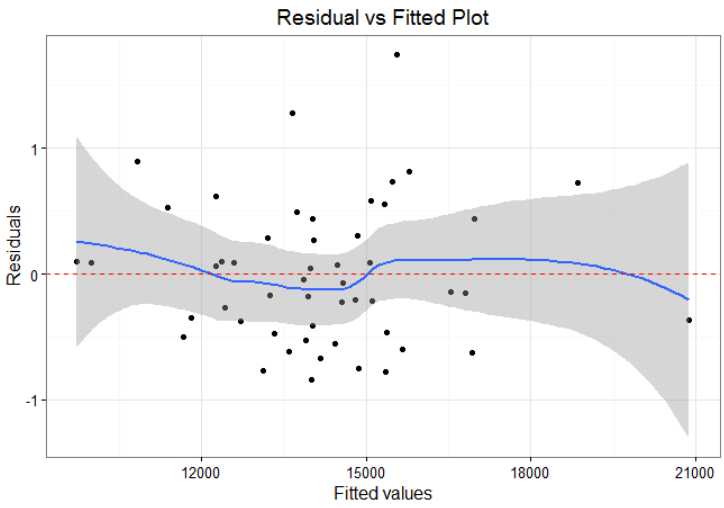


Τέλος μη ικανοποιητικό κρίνεται και το μοντέλο που περιλαμβάνει τον μέσο αριθμό μερών εργασίας και την παραγωγή κάθε κατηγορίας αγροτικού προιόντος:

```
##                                                                  Call:
## glm(formula = ΕΚΜΕΤΑΛΛΕΥΣΕΙΣ ~ Μέρες εργασίας κατ. 1 + Μέρες εργασία
ς      κατ.      2      +      Μέρες      εργασίας      κατ.      3      +
##         Μέρες  εργασίας  κατ.  4,  family  =  "poisson",  data  =  df2)
##
##                             Deviance                             Residuals:
##          Min               1Q      Median               3Q           Max
##   -123.35         -58.36           -8.79           42.10         178.56
##
##                                                          Coefficients:
##                             Estimate  Std.  Error  z  value  Pr(>|z|)
##  (Intercept)     9.58e+00        7.33e-03   1306.04        <2e-16   ***
## Μέρες εργασίας κατ. 1    4.46e-03    4.26e-05   104.52    <2e-16 ***
```

```
## Μέρες εργασίας κατ. 2   -1.64e-03    3.12e-05   -52.75    <2e-16 ***
## Μέρες εργασίας κατ. 3   -1.93e-03    2.06e-04    -9.37    <2e-16 ***
## Μέρες εργασίας κατ. 4   -2.08e-02    4.17e-04   -49.90    <2e-16 ***
##                                                           ---
## Signif. codes:   0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##   (Dispersion parameter for poisson family taken to be 1)
##
##          Null deviance: 239137    on 50    degrees of freedom
##   Residual deviance:  224955     on 46     degrees of freedom
##                              AIC:                          225537
##
## Number of Fisher Scoring iterations: 5
```



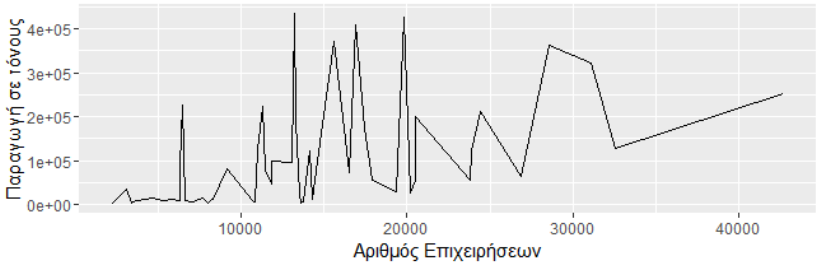


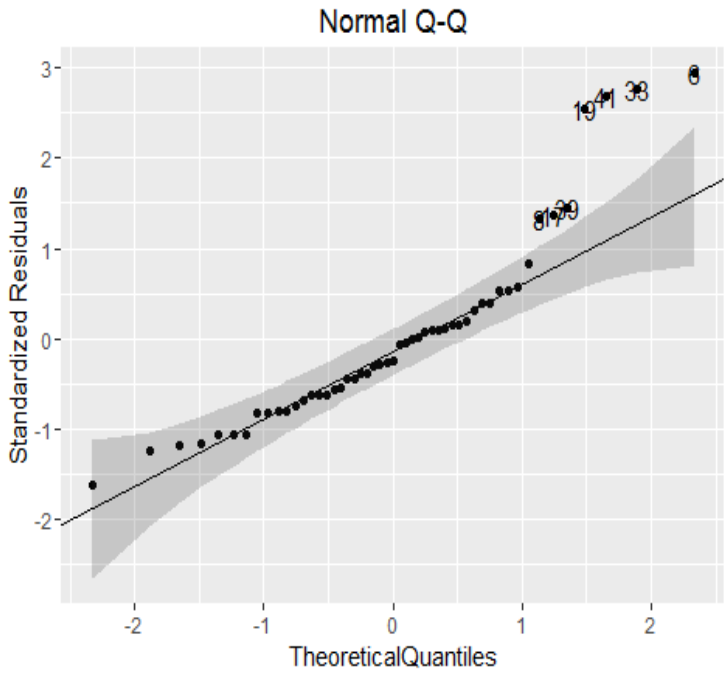
Παλινδρόμηση για την πρόβλεψη της συνολικής παραγωγής σε σχέση με τις μέσες εργασίας

Στην ενότητα αυτή θα προσπαθήσουμε να μοντελοποιήσουμε την συνολική παραγωγή κάθε νομού σε σχέση με τις μέσες μέρες εργασίας, τον αριθμό κάθε κατηγορίας εργαζομένων, το σύνολο των εκμεταλλεύσεων και τον πληθυσμό κάθε νομού. Επειδή έχουμε πολύ διαφορετικές κλίμακες μεγεθών, θα χρησιμοποιήσουμε τις μεταβλητές κανονικοποιήμενες.

```
##     lm(formula = Production ~ ., data = df2a)
##
##                                                 Residuals:
##            Min              1Q  Median              3Q             Max
##      -1.278      -0.545      -0.200           0.269           2.262
##
##                                                 Coefficients:
##                              Estimate  Std. Error  t value  Pr(>|t|)
##  (Intercept)            -3.33e-17      1.26e-01       0.00      1.0000
## ΕΚΜΕΤΑΛΛΕΎΣΕΙΣ             4.53e-01    1.32e-01     3.43    0.0014 **
##  Faily_Members           2.84e-02      2.47e-01       0.12      0.9089
## Μέρες εργασίας κατ. 1          2.01e-01    1.41e-01     1.42    0.1626
##  Εργαζόμενοι κατ.  2     -1.35e-01     1.76e-01     -0.77     0.4478
## Μέρες εργασίας κατ. 2         -3.51e-02    1.52e-01    -0.23    0.8185
##  Εργαζόμενοι κατ.  3    1.18e-01      2.17e-01       0.54      0.5909
## Μέρες εργασίας κατ. 3         -1.44e-01    1.48e-01    -0.97    0.3362
## Εργαζόμενοι κατ. 4             -2.39e-01    1.85e-01    -1.29    0.2037
## Μέρες εργασίας κατ. 4         -1.52e-01    1.63e-01    -0.93    0.3568
##                                                                    ---
## Signif. codes:   0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##  Residual standard error: 0.897 on 41 degrees of freedom
##  Multiple R-squared:     0.341,     Adjusted R-squared:     0.196
## F-statistic: 2.35 on 9 and 41 DF,  p-value: 0.0303
```

Το μοντέλο δεν είναι καλά ορισμένο όπως είναι φανερό τόσο από το summary όσο και από το διαγνωστικά γραφήματα. Ο μόνος παράγοντας που προκύπτει στατιστικά σημαντικός σε σχέση με τη συνολική παραγωγή είναι ο αριθμός των επιχειρήσεων. Η σχέση επιχειρήσεων και παραγωγής φαίνεται στο παρακάτω γράφημα:





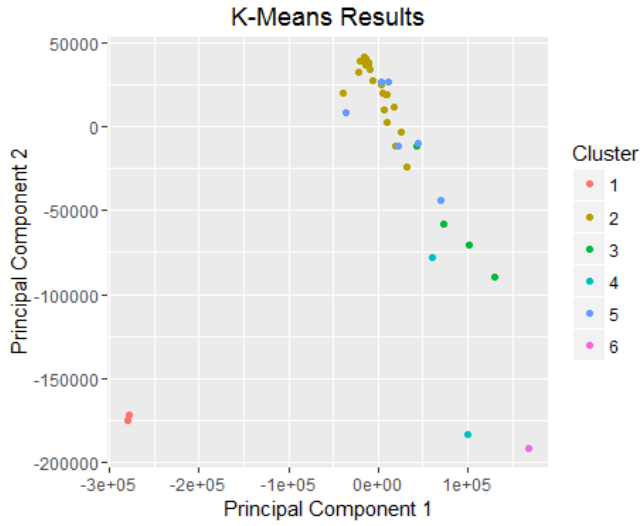
Δοκιμάζουμε ακόμα ένα μοντέλο χωρίς την μεταβλητή με τον αριθμό των επιχειρήσεων, χωρίς κάποιο θετικό αποτέλεσμα.

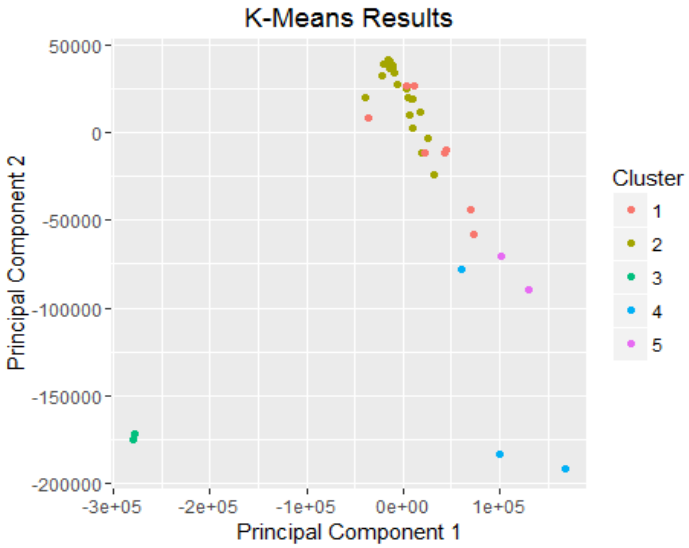
```
Call:
##    lm(formula    =    Production    ~    .,    data    =    df3)
##
##                                                          Residuals:
##           Min              1Q   Median              3Q           Max
##      -1.510      -0.642      -0.307           0.429           2.525
##
##                                                       Coefficients:
##                                Estimate  Std.  Error  t  value  Pr(>|t|)
##  (Intercept)           -2.70e-17     1.41e-01       0.00       1.000
##  Faily_Members        -1.02e-02      2.77e-01      -0.04        0.971
## Μέρες εργασίας κατ. 1         3.08e-01    1.55e-01     1.99     0.053 .
##  Εργαζόμενοι  κατ.  2     -5.56e-02     1.96e-01     -0.28       0.778
## Μέρες εργασίας κατ. 2         -9.38e-02    1.69e-01    -0.55     0.582
##  Εργαζόμενοι  κατ.  3    1.50e-01      2.43e-01       0.62       0.541
## Μέρες εργασίας κατ. 3         -1.78e-01    1.65e-01    -1.08     0.287
## Εργαζόμενοι κατ. 4             -2.63e-01    2.07e-01    -1.27     0.211
## Μέρες εργασίας κατ. 4         -1.88e-01    1.83e-01    -1.03     0.310
##                                                                    ---
##  Signif.  codes:   0  '***'  0.001  '**'  0.01  '*'  0.05  '.'  0.1  ' '  1
##
##    Residual   standard   error:   1   on   42   degrees   of   freedom
##   Multiple  R-squared:     0.152,     Adjusted  R-squared:     -0.00998
## F-statistic: 0.938 on 8 and 42 DF,  p-value: 0.496
```

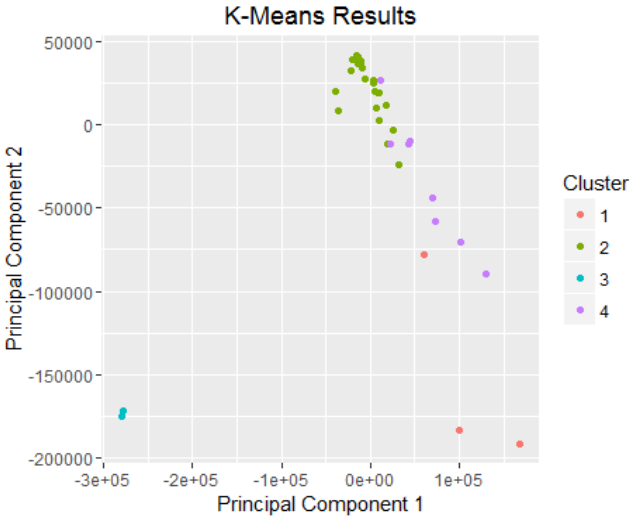
Εάν είχαμε το σύνολο της πραγματικής παραγωγής του κάθε νομού ένα τέτοιο μοντέλο θα μας έδειχνε το παράδοξο αποτέλεσμα ότι η γεωργική παραγωγή δεν εξαρτάται από το πόσοι άνθρωποι και για πόσο χρόνο εργάζονται για αυτή. Φυσικά κάτι τέτοιο θα ήταν μια εντελώς εσφαλμένη κρίση.

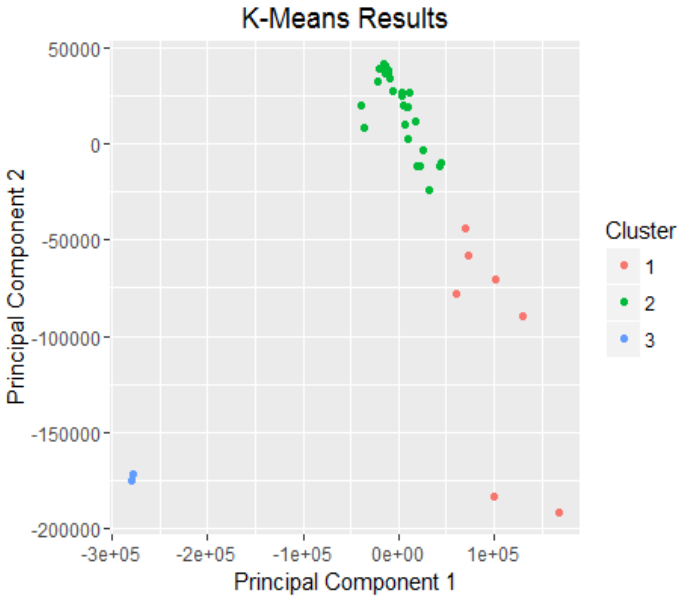
Ομαδοποίηση νομών σε σχέση με την παραγωγή κάθε προϊόντος

Θα εξετάσουμε ομαδοποιήσεις στα δεδομένα σε σχέση με την παραγωγή κάθε νομού. Για την ενότητα αυτή θα εργαστούμε με τον αλγόριθμο k-means.




K-Means Results
50000
0
-50000
-100000
-150000
-200000
Principal Component 2
-3e+05
-2e+05
-1e+05
0e+00
1e+05
Principal Component 1
Cluster
1
2
3
4
5


K-Means Results
50000
0
-50000
-100000
-150000
-200000
Principal Component 2
-3e+05
-2e+05
-1e+05
0e+00
1e+05
Principal Component 1
Cluster
1
2
3
4



Ο αλγόριθμος διαχωρίζει τους νομούς στα παρακάτω 6 Clusters:

Cluster 1

ΝΟΜΟΣ ΗΜΑΘΙΑΣ, ΝΟΜΟΣ ΠΕΛΛΗΣ

Cluster 2

ΝΟΜΟΣ ΑΤΤΙΚΗΣ,ΝΟΜΟΣ ΑΙΤΩΛΙΑΣ ΚΑΙ ΑΚΑΡΝΑΝΙΑΣ, ΝΟΜΟΣ ΑΡΚΑΔΙΑΣ, ΝΟΜΟΣ ΑΧΑΙΑΣ,ΝΟΜΟΣ ΒΟΙΩΤΙΑΣ,ΝΟΜΟΣ ΓΡΕΒΕΝΩΝ,ΝΟΜΟΣ ΔΡΑΜΑΣ, ΝΟΜΟΣ ΔΩΔΕΚΑΝΗΣΟΥ,ΝΟΜΟΣ ΕΒΡΟΥ,ΝΟΜΟΣ ΕΥΒΟΙΑΣ,ΝΟΜΟΣ ΕΥΡΥΤΑΝΙΑΣ,ΝΟΜΟΣ ΘΕΣΠΡΩΤΙΑΣ,ΝΟΜΟΣ ΘΕΣΣΑΛΟΝΙΚΗΣ,ΝΟΜΟΣ ΙΩΑΝΝΙΝΩΝ,ΝΟΜΟΣ ΚΑΡΔΙΤΣΗΣ,ΝΟΜΟΣ ΚΑΣΤΟΡΙΑΣ,ΝΟΜΟΣ ΚΕΦΑΛΛΗΝΙΑΣ,ΝΟΜΟΣ ΚΙΛΚΙΣ,ΝΟΜΟΣ ΚΟΖΑΝΗΣ,ΝΟΜΟΣ ΚΥΚΛΑΔΩΝ,ΝΟΜΟΣ ΛΕΣΒΟΥ,ΝΟΜΟΣ ΛΕΥΚΑΔΟΣ,ΝΟΜΟΣ ΜΑΓΝΗΣΙΑΣ,ΝΟΜΟΣ ΞΑΝΘΗΣ,ΝΟΜΟΣ ΠΙΕΡΙΑΣ,ΝΟΜΟΣ ΠΡΕΒΕΖΗΣ,ΝΟΜΟΣ ΡΕΘΥΜΝΗΣ,ΝΟΜΟΣ ΡΟΔΟΠΗΣ,ΝΟΜΟΣ ΣΑΜΟΥ,ΝΟΜΟΣ ΣΕΡΡΩΝ,ΝΟΜΟΣ ΤΡΙΚΑΛΩΝ,ΝΟΜΟΣ ΦΘΙΩΤΙΔΟΣ,ΝΟΜΟΣ ΦΛΩΡΙΝΗΣ,ΝΟΜΟΣ ΦΩΚΙΔΟΣ,ΝΟΜΟΣ ΧΑΛΚΙΔΙΚΗΣ,ΝΟΜΟΣ ΧΙΟΥ

Cluster 3

ΝΟΜΟΣ ΗΡΑΚΛΕΙΟΥ, ΝΟΜΟΣ ΚΕΡΚΥΡΑΣ,ΝΟΜΟΣ ΜΕΣΣΗΝΙΑΣ, ΝΟΜΟΣ ΧΑΝΙΩΝ

Cluster 4

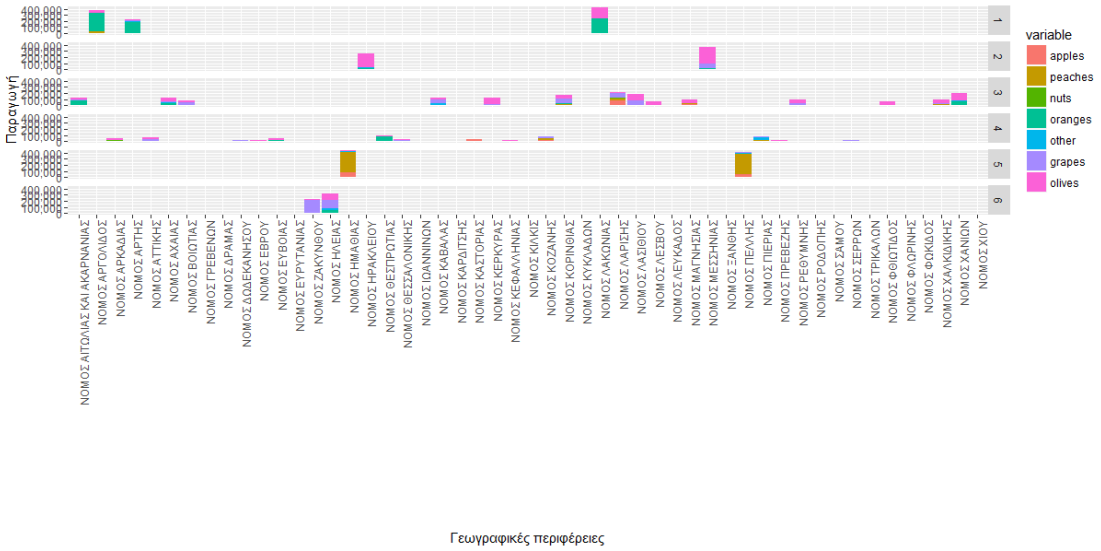
ΝΟΜΟΣ ΑΡΓΟΛΙΔΟΣ, ΝΟΜΟΣ ΑΡΤΗΣ

Cluster 5

ΝΟΜΟΣ ΖΑΚΥΝΘΟΥ,ΝΟΜΟΣ ΗΛΕΙΑΣ,ΝΟΜΟΣ ΚΑΒΑΛΑΣ,ΝΟΜΟΣ ΚΟΡΙΝΘΙΑΣ, ΝΟΜΟΣ ΛΑΡΙΣΗΣ,ΝΟΜΟΣ ΛΑΣΙΘΙΟΥ

Cluster 6

ΝΟΜΟΣ ΛΑΚΩΝΙΑΣ

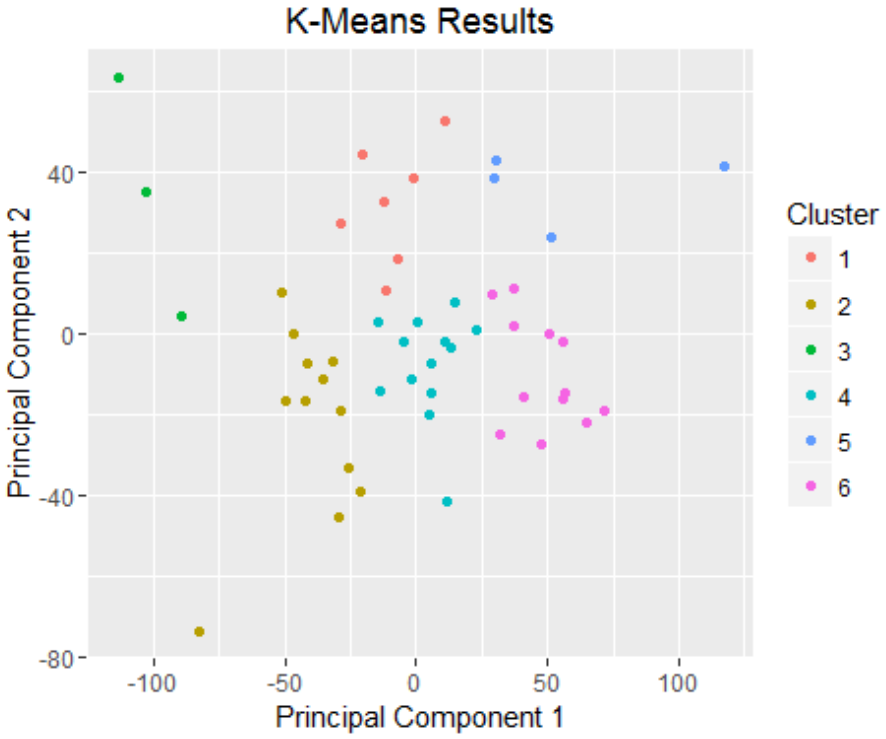


Μπορούμε να ερμηνεύσουμε ποιοτικά το αποτέλεσμα από το παραπάνω γράφημα, με τον αλγόριθμο να χωρίζει τους νομούς:

- Σε αυτούς που παράγουν σε μεγάλες ποσότητες (>200 tn) πορτοκαλιά και παρόμοια φρούτα - Cluster 1
- Σε αυτούς που παράγουν σε μεγάλες ποσότητες και κατά κύριο λόγο ελιές Cluster 2
- Σε αυτούς που παράγουν σε μικρή (50-60 tn) ποσότητες και κατά κύριο λόγο ελιές Cluster 3
- Στους νομούς με πολύ μικρή αγροτική παραγωγή (συνολικά < 100tn) Cluster 4
- Σε αυτούς που παράγουν σε μεγάλες ποσότητες ροδάκινα Cluster 5
- Σε αυτούς που παράγουν σε μεσαίες (<200 tn & >80 tn) ποσότητες σταφύλια Cluster 6

Το clustering αυτό εξηγεί το 85.6% του συνολικού variability των δεδομένων.

Ομαδοποίηση νομών σε σχέση με τις ημέρες εργασίας



Cluster 1

ΝΟΜΟΣ ΗΜΑΘΙΑΣ, ΝΟΜΟΣ ΠΕΛΛΗΣ

Cluster 2

ΝΟΜΟΣ ΑΤΤΙΚΗΣ, ΝΟΜΟΣ ΑΙΤΩΛΙΑΣ ΚΑΙ ΑΚΑΡΝΑΝΙΑΣ, ΝΟΜΟΣ ΑΡΚΑΔΙΑΣ,ΝΟΜΟΣ ΑΧΑΙΑΣ,ΝΟΜΟΣ ΒΟΙΩΤΙΑΣ,ΝΟΜΟΣ ΓΡΕΒΕΝΩΝ,ΝΟΜΟΣ ΔΡΑΜΑΣ,ΝΟΜΟΣ ΔΩΔΕΚΑΝΗΣΟΥ,ΝΟΜΟΣ ΕΒΡΟΥ,ΝΟΜΟΣ ΕΥΒΟΙΑΣ,ΝΟΜΟΣ ΕΥΡΥΤΑΝΙΑΣ,ΝΟΜΟΣ ΘΕΣΠΡΩΤΙΑΣ,ΝΟΜΟΣ ΘΕΣΣΑΛΟΝΙΚΗΣ,ΝΟΜΟΣ ΙΩΑΝΝΙΝΩΝ,ΝΟΜΟΣ ΚΑΡΔΙΤΣΗΣ,ΝΟΜΟΣ ΚΑΣΤΟΡΙΑΣ, ΝΟΜΟΣ ΚΕΦΑΛΛΗΝΙΑΣ,ΝΟΜΟΣ ΚΙΛΚΙΣ,ΝΟΜΟΣ ΚΟΖΑΝΗΣ,ΝΟΜΟΣ ΚΥΚΛΑΔΩΝ,ΝΟΜΟΣ ΛΕΣΒΟΥ,ΝΟΜΟΣ ΛΕΥΚΑΔΟΣ,ΝΟΜΟΣ ΜΑΓΝΗΣΙΑΣ, ΝΟΜΟΣ ΞΑΝΘΗΣΝΟΜΟΣ ΠΙΕΡΙΑΣ,ΝΟΜΟΣ ΠΡΕΒΕΖΗΣ,ΝΟΜΟΣ ΡΕΘΥΜΝΗΣ,ΝΟΜΟΣ ΡΟΔΟΠΗΣ,ΝΟΜΟΣ ΣΑΜΟΥ,ΝΟΜΟΣ ΣΕΡΡΩΝ,ΝΟΜΟΣ ΤΡΙΚΑΛΩΝ, ΝΟΜΟΣ ΦΘΙΩΤΙΔΟΣ,ΝΟΜΟΣ ΦΛΩΡΙΝΗΣ,ΝΟΜΟΣ ΦΩΚΙΔΟΣ,ΝΟΜΟΣ ΧΑΛΚΙΔΙΚΗΣ,,ΝΟΜΟΣ ΧΙΟΥ

Cluster 3

ΝΟΜΟΣ ΗΡΑΚΛΕΙΟΥ, ΝΟΜΟΣ ΚΕΡΚΥΡΑΣ,ΝΟΜΟΣ ΜΕΣΣΗΝΙΑΣ, ΝΟΜΟΣ ΧΑΝΙΩΝ

Cluster 4

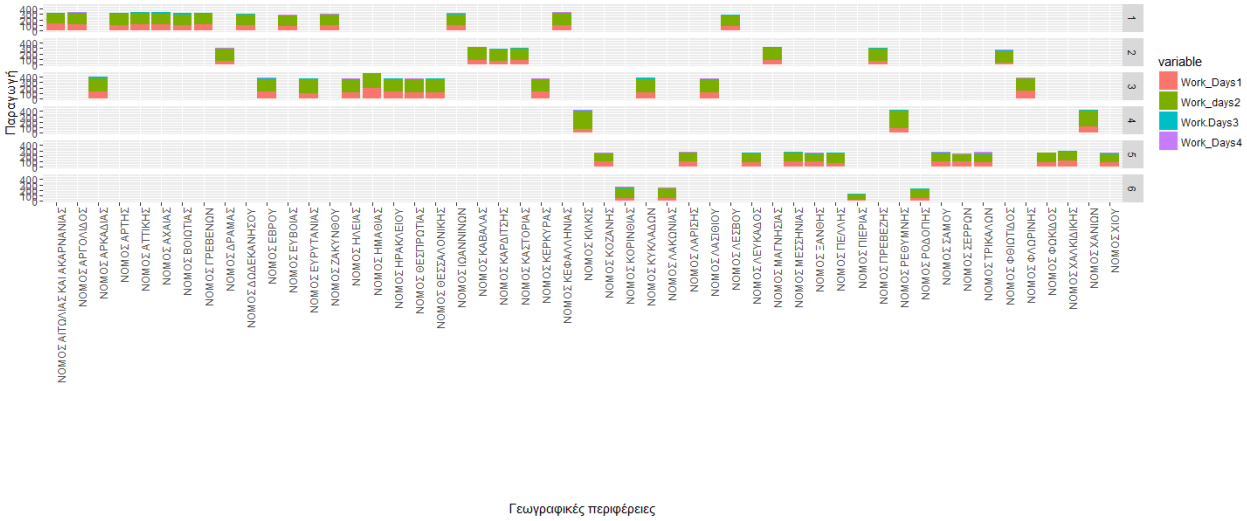
ΝΟΜΟΣ ΑΡΓΟΛΙΔΟΣ, ΝΟΜΟΣ ΑΡΤΗΣ

Cluster 5

ΝΟΜΟΣ ΖΑΚΥΝΘΟΥ,ΝΟΜΟΣ ΗΛΕΙΑΣ,ΝΟΜΟΣ ΚΑΒΑΛΑΣ,ΝΟΜΟΣ ΚΟΡΙΝΘΙΑΣ, ΝΟΜΟΣ ΛΑΡΙΣΗΣ,ΝΟΜΟΣ ΛΑΣΙΘΙΟΥ

Cluster 6

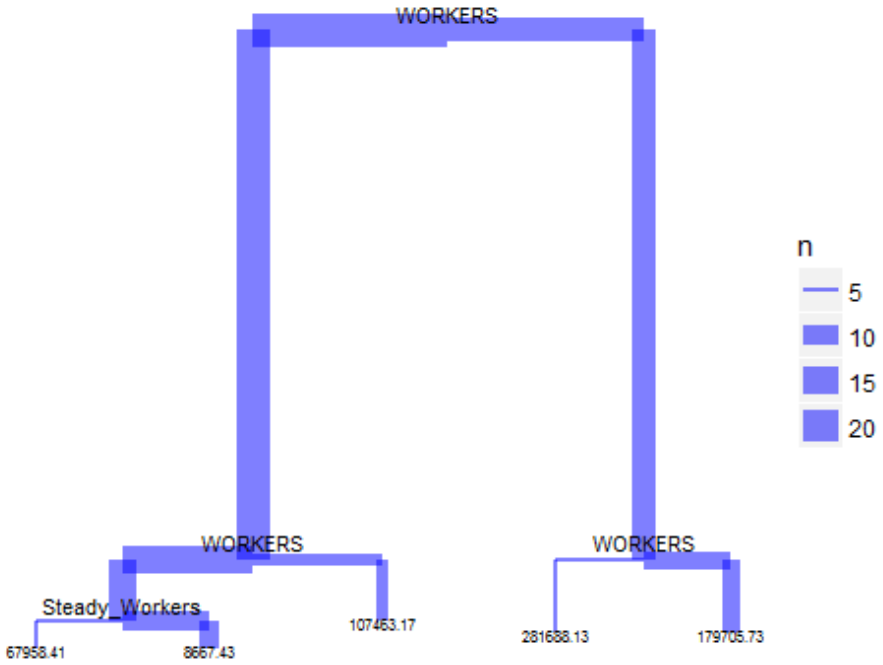
ΝΟΜΟΣ ΛΑΚΩΝΙΑΣ

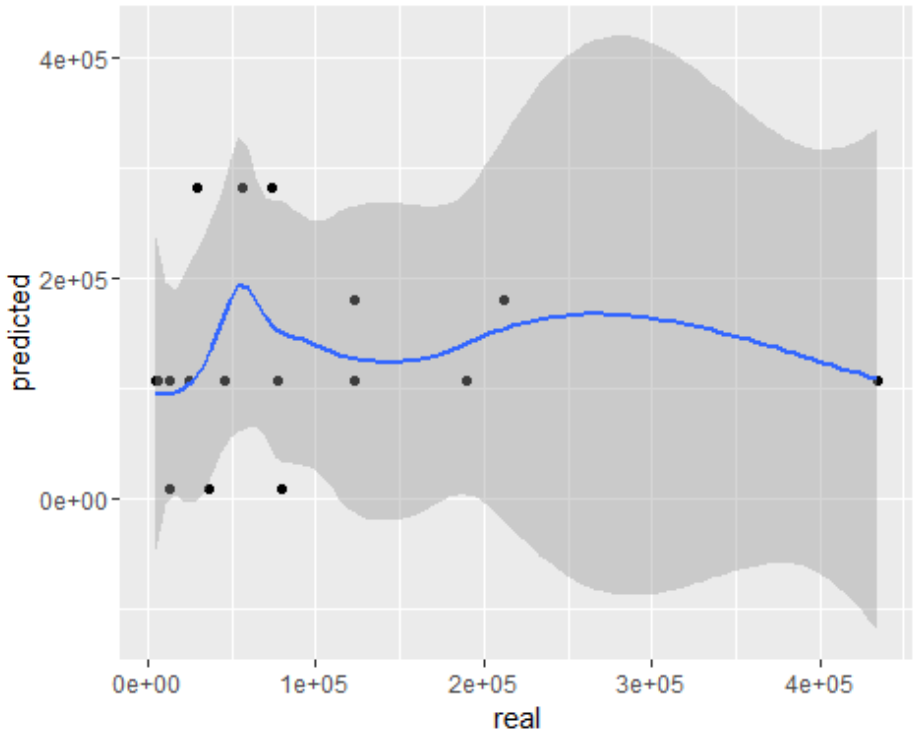


Στην περίπτωση αυτή όπου το το clustering εξηγεί το 80.3% του συνολικού variability των δεδομένων. Δεν είναι τόσο εύκολο να βγάλουμε συμπεράσματα όπως παραπάνω αλλά παρατηρούμε ότι οι δύο ομαδοποιήσεις δεν ταυτίζονται.

Δεντρά regression για την μοντελοποίηση της συνολικής παραγωγής κάθε νομού

Εφόσον στην προσπάθεια μας να μοντελοποιήσουμε την συνολική παραγωγή κάθε νομού σε σχέση με τις μέσες μέρες εργασίας, τον αριθμό κάθε κατηγορίας εργαζομένων κτλ. με γραμμική παλινδρόμηση δεν είχαμε καλά αποτελέσματα, θα δοκιμάσουμε το ίδιο με δέντρα παλινδρόμησης. Στην ενότητα αυτή, παρόλο που το dataset είναι πολύ μικρό, για τις ανάγκες της άσκησης θα χρησιμοποιήσουμε τα 2/3 του μοντέλου για εκπαίδευση και το 1/3 για επαλήθευση του μοντέλου.





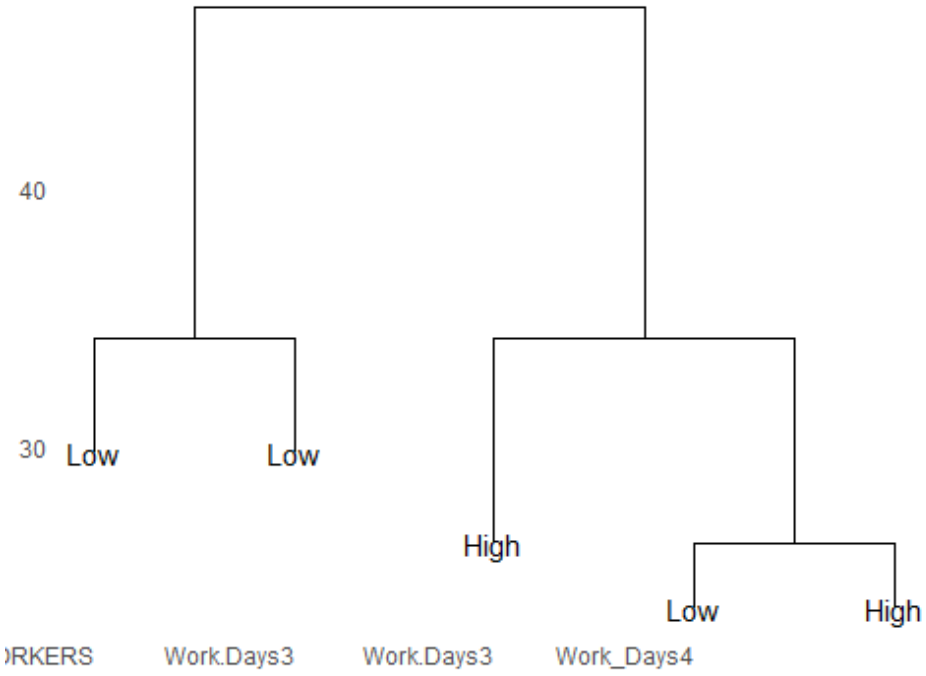

```
## [1] 1.9e+10
```

Το από το δέντρο φαίνεται ότι οι μεταβλητές που επηρεάζουν την παραγωγή είναι ο αριθμός των εκμεταλλεύσεων καθώς και ο αριθμός των σταθερών εργαζομένων.

Το μοντέλο φαίνεται να έχει πολύ κακή προσαρμογή στα πραγματικά δεδομένα και μάλιστα έχουμε πάρα πολύ μεγάλο μέσο τετραγωνικό σφάλμα: R^2= 1.9 * e^10

Classification της παραγωγής σε χαμηλή-ψηλή

Για τις ανάγκες της άσκησης θα εφαρμόσουμε δύο τεχνικές ταξινόμησης, με δένδρα και SVM. Για να γίνει αυτό μετατρέπουμε την μεταβλητή σχετικά με την συνολική παραγωγή σε μια νέα μεταβλητή, κατηγορική με 2 τιμές "Χαμηλή" παραγωγή και "Ψηλή" παραγωγή. Ο διαχωρισμός έγινε αυθαίρετα, έτσι ώστε το σετ να είναι χωρισμένο στους 26 νομούς με την χαμηλότερη και 25 με την υψηλότερη.



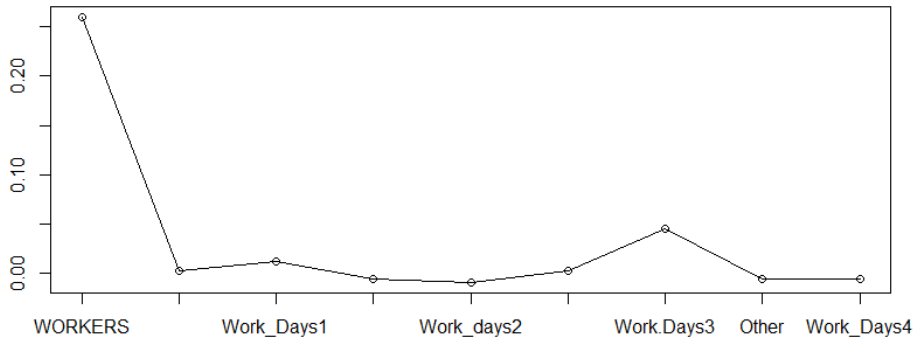
	Low	High
Low	6	3
High	4	4

Όπως φαίνεται στο confusion Matrix, το μοντέλο ταξινόμησης με δέντρο δεν δίνει καθόλου καλά αποτελέσματα.

Σχετικά καλύτερα αποτελέσματα παίρνουμε δοκιμάζοντας ένα ensemble 1000 δέντρων με την τεχνική Random Forest:

```
    Low High
Low   5    1
High  5    6

 Accuracy : 0.647
  95% CI : (0.383, 0.858)
 No Information Rate : 0.588
 P-Value [Acc > NIR] : 0.409
```

Σημαντικότερο ρόλο για την παραγωγή φαίνεται να παίζει ο αριθμός των εκμεταλλεύσεων (Workers) και οι ημέρες εργασίας των εποχιακών εργατών (Work.Days3).

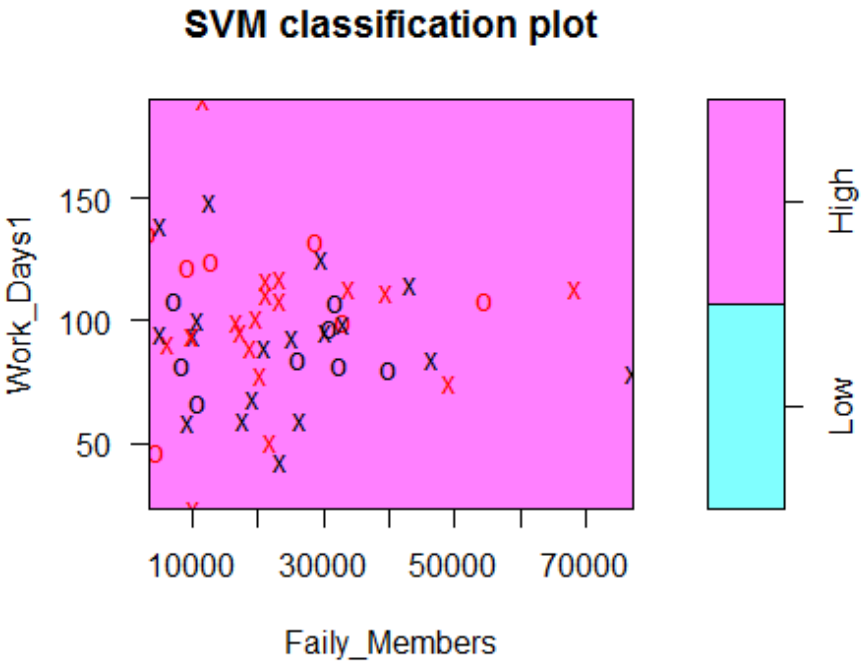


Δοκιμάζοντας SVM αρχικά θα ψάξουμε για τις καλύτερες παραμέτρους, για radial και linear πυρήνες με διαφορετικό κόστος ανάμεσα στα (0.001,0.01,0.1,0.5,1,1.5,10,20,30,40,50,60,80,40,90,100,150,200).

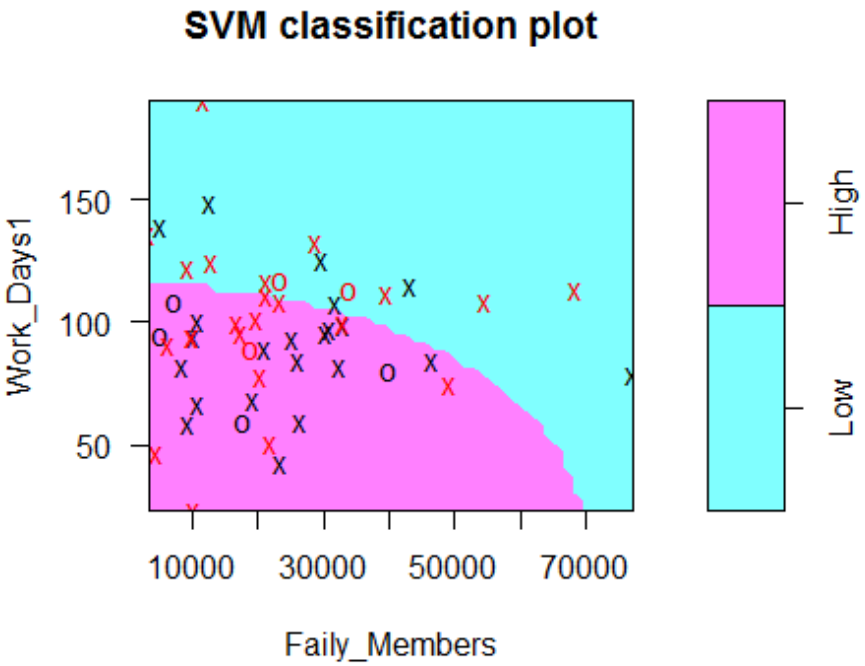
```
##
##           Parameter            tuning           of           'svm':
##
##     -     sampling     method:     10-fold     cross     validation
##
##                 -                best                 parameters:
##                                                              cost
##                                                                50
##
## - best performance: 0.433

##
##           Parameter            tuning           of           'svm':
```

```
##
##     -     sampling     method:     10-fold     cross     validation
##
##                  -                 best                 parameters:
##                                                                cost
##                                                                  20
##
## - best performance: 0.43
```

Εμπιστευόμενοι τα παραπάνω καταλήγουμε στο γραμμικό πυρήνα με κόστος 50.Απο το γράφημα όμως βλέπουμε ότι έχουμε overfiting.



Για τον radial πυρήνα, τα αποτελέσματα δεν είναι πολύ καλύτερα:



Συμπεράσματα

Λόγω της φύσης του dataset, το οποίο ήταν ελλιπές από την άποψη της καταγραφής της γεωργικής παραγωγής καθώς περιείχε μόνο προϊόντα από δέντρα, αλλά και λόγω του ότι αφορούσε μόνο μια χρονιά, δεν μπορούμε να ισχυριστούμε πως τα όποια αποτελέσματα μας είναι σημαντικά ή αξιόπιστα. Παρόλα αυτά για τις ανάγκες τις άσκησης θα τα παραθέσουμε συνοπτικά:

- Οι γεωργικές επιχειρήσεις στην Ελλάδα είναι κατά βάσει οικογενειακές και οι εκάστοτε ανάγκες για παραπάνω εργατικά χέρια καλύπτονται ως επί το πλείστο από εποχιακούς εργάτες.
- Τα περισσότερα αγροτικά προϊόντα στην Ελλάδα παράγονται στις περιφέρειες της Πελοποννήσου, της Κεντρικής Μακεδονίας, της Κρήτης και της Δυτικής Μακεδονίας

Τέλος μπορούμε να διαχωρίσουμε ποιοτικά τους διάφορους νομούς ανάλογα με την παραγωγή τους σε κατηγορίες:

- Σε αυτούς που παράγουν σε μεγάλες ποσότητες (>200 tn) πορτοκαλιά και παρόμοια φρούτα
- Σε αυτούς που παράγουν σε μεγάλες ποσότητες και κατά κύριο λόγο ελιές
- Σε αυτούς που παράγουν σε μικρή (50-60 tn) ποσότητες και κατά κύριο λόγο ελιές
- Στους νομούς με πολύ μικρή αγροτική παραγωγή (συνολικά < 100tn)
- Σε αυτούς που παράγουν σε μεγάλες ποσότητες ροδάκινα
- Σε αυτούς που παράγουν σε μεσαίες (<200 tn & >80 tn) ποσότητες σταφύλια