MSc in Business Analytics

Business Analytics I

Main Assignment

Τακλάκογλου Χιδίρογλου Αργύριος
AM: f2822114
Dataset 14

ΠΕΡΙΕΧΟΜΕΝΑ

Εισαγωγή

Τα συστήματα ενοικίασης ποδηλάτων είναι η νέα γενιά ενοικίασης ποδηλάτων όπου όλη η διαδικασία από την εγγραφή στο σύστημα, την ενοικίαση του ποδηλάτου και την επιστροφή του, γίνεται αυτόματα. Με αυτά τα συστήματα, ο χρήστης μπορεί εύκολα να νοικιάσει ένα ποδήλατο από μια συγκεκριμένη θέση και να το επιστρέψει σε διαφορετική τοποθεσία. Σήμερα υπάρχουν περίπου 500 συστήματα bike sharing σε όλο τον κόσμο τα όποια συνολικά κατέχουν παραπάνω από 500 χιλιάδες ποδήλατα. Λόγω της κυκλοφοριακής συμφόρησης αλλά και των περιβαλλοντικών προβλημάτων, η χρήση αυτοκινήτου δεν είναι αποτελεσματική για μικρές αποστάσεις. Η χρήση του ποδηλάτου ως μέσο μεταφοράς έχει αυξηθεί τα τελευταία χρόνια. Τα συστήματα ενοικίασης ποδηλάτων είναι μία καλή λύση για μετακίνηση εντός πόλης διότι δεν βλάπτουν το περιβάλλον, τα έξοδα καυσίμου είναι περισσότερα από το κόστος ενοικίασης ενός ποδηλάτου, οι ποδηλάτες ασκούνται και δεν χρειάζεται να σκέφτονται το παρκινγκ. Όπως είναι φυσικό, λόγω των αυτοματοποιημένων και πολυάριθμων καθημερινών ενοικιάσεων, παράγονται πολλά δεδομένα στα συστήματα αυτά. Τα δεδομένα αυτά μπορούν να χρησιμοποιηθούν για έρευνα καθώς το σημείο αναχώρησης και άφιξης δεν είναι προκαθορισμένο. Αυτό το χαρακτηριστικό μετατρέπει το σύστημα κοινής χρήσης ποδηλάτων σε ένα εικονικό δίκτυο αισθητήρων που μπορεί να χρησιμοποιηθεί για ανίχνευση κινητικότητα στην πόλη. Σε αυτή την εργασία θα πραγματοποιήσω διερευνητική ανάλυση στα δεδομένα προκειμένου να βρω σχέσεις μεταξύ των μεταβλητών όπως για παράδειγμα ποιες μεταβλητές επηρεάζουν το σύνολο των ενοικιάσεων ποδηλάτου και πως τις επηρεάζουν καθώς και να επιχειρήσω να φτιάξω ένα μοντέλο πρόβλεψης του συνόλου των ενοικιάσεων ποδηλάτου για να προβλέψουμε την ζήτηση.

Το σετ δεδομένων αποτελείται από δεδομένα των ετών 2011 και 2012 της Capital Bikeshare system, Washington D.C., USA στα οποία προστέθηκαν οι αντίστοιχες καιρικές και εποχικές πληροφορίες από το http://www.freemeteo.com. Αποτελείται από 18 μεταβλητές και 1500 παρατηρήσεις. Πιο συγκεκριμένα οι μεταβλητές μας είναι οι εξής:

1) Η μεταβλητή X και η μεταβλητή instant που περιέχουν τους κωδικούς των ενοικιάσεων.
2) Η μεταβλητή dteday που περιέχει την ημερομηνία της κάθε ενοικίασης.
3) Η μεταβλητή season που περιέχει την εποχή που πραγματοποιήθηκε η κάθε ενοικίαση. Όπου 1 = Χειμώνας , 2 = Άνοιξη, 3 = Καλοκαίρι και 4 = Φθινόπωρο
4) Η μεταβλητή yr που περιέχει το έτος που πραγματοποιήθηκε η κάθε ενοικίαση. Όπου 0 = 2011 και 1 = 2012
5) Η μεταβλητή Mnth που περιέχει τον μήνα που πραγματοποιήθηκε η κάθε ενοικίαση.
6) Η μεταβλητή hr που περιέχει την ώρα που πραγματοποιήθηκε η κάθε ενοικίαση.
7) Η μεταβλητη holiday που περιέχει το αν ήταν ημέρα διακοπών ή όχι η ημέρα που πραγματοποιήθηκε η κάθε ενοικίαση. Όπου 0 = δεν ήταν ημέρα διακοπών και 1 = ήταν ημέρα διακοπών.
8) Η μεταβλητή weekday που περιέχει την ήμερα της εβδομάδας που πραγματοποιήθηκε η κάθε ενοικίαση.
9) Η μεταβλητή workingday που περιέχει το αν ήταν ημέρα διακοπών ή ημέρα αργίας ή όχι η ημέρα που πραγματοποιήθηκε η ενοικίαση. Όπου 0 = ήταν ημέρα διακοπών ή ημέρα αργίας και 1 = ήταν μία εργάσιμη ημέρα
10) Η μεταβλητή weathersit που περιέχει τις καιρικές συνθήκες που επικρατούσαν την ημέρα της ενοικίασης. Όπου:
    - 1 = Good = Clear, Few clouds, Partly cloudy,
    - 2 = Medium = Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist

- 3 = Bad = Light Snow, Light Rain + Thunderstorm + Scattered Clouds, Light Rain + Scattered clouds
- 4 = Very Bad = Heavy Rain + Ice Pallets + Thunderstorm + Mist Snow + Fog

11) Η μεταβλητή temp που περιέχει την θερμοκρασία που επικρατούσε την ώρα της ενοικίασης.
12) Η μεταβλητή atemp που περιέχει το πως αισθανόσουν την θερμοκρασία που επικρατούσε την ώρα της ενοικίασης. Το πόσους βαθμούς δηλαδή πίστευαν ότι επικρατούσαν την στιγμή της ενοικίασης.
13) Η μεταβλητή hum που περιέχει την υγρασία που επικρατούσε την στιγμή της ενοικίασης.
14) Η μεταβλητή windspeed που περιέχει την ταχύτητα του αέρα που επικρατούσε την στιγμή της ενοικίασης.
15) Η μεταβλητή casual που περιέχει το σύνολο των ενοικιάσεων κάθε ώρα από περιστασιακούς πελάτες.
16) Η μεταβλητή registered που περιέχει το σύνολο των ενοικιάσεων κάθε ώρα από εγγεγραμμένους πελάτες.
17) Η μεταβλητή cnt που περιέχει το σύνολο των ενοικιάσεων κάθε ώρα και από casual αλλά και από registered πελάτες.

Κάνοντας περιγραφή των δεδομένων παρατήρησα ότι δεν υπήρχαν κενές τιμές(missing values) στο σετ δεδομένων μου καθώς και οτι πολλές μεταβλητές δεν ήταν σωστά ορισμένες. Γι' αυτό άλλαξα τον τύπο των μεταβλητών. Πιο συγκεκριμένα:

- το dteday από μεταβλητές χαρακτήρων σε date
- το season, το yr , το mnth, το hr, το holiday, το weekday, το workingday και το weathersit από μεταβλητές χαρακτήρων(character) σε κατηγορικές(factors)
- το temp, το atemp, το hum, το windspeed, και το cnt από μεταβλητές ακαίρων(integer) σε αριθμητικές(numeric). Επίσης πολλαπλασίασα το temp με το 41 , το atemp με το 50, το hum με το 100 και το windspeed με το 67

Έπειτα αφαίρεσα τις μεταβλητές X, instant, dteday, casual, registered από το σετ δεδομένων, καθώς το X και το instant δεν παρέχουν χρήσιμα δεδομένα για την ανάλυση μας. Το dteday δεν μας χρειάζεται καθώς παρέχονται δεδομένα για κάθε έτος, κάθε μήνα, κάθε ημέρα και κάθε ώρα των ενοικιάσεων. Τέλος, αφαίρεσα τις μεταβλητές casual και registered, διότι αθροίζοντάς τις παρατηρήσεις τους καταλήγουμε στις παρατηρήσεις της μεταβλητής cnt.

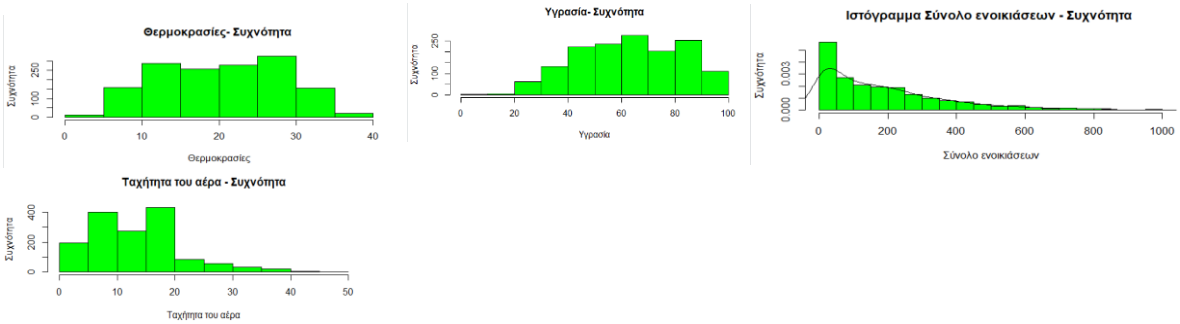
Περιγραφική ανάλυση και διερευνητική ανάλυση δεδομένων

Αρχικά, πραγματοποίησα περιγραφική ανάλυση των δεδομένων, ήλεγξα για κενές τιμές και άλλαξα των τύπο των δεδομένων σε κάποιες μεταβλητές. Έπειτα έκανα ιστόγραμμα για κάθε ποσοτική μεταβλητή και boxplots για τις κατηγορικές μεταβλητές.

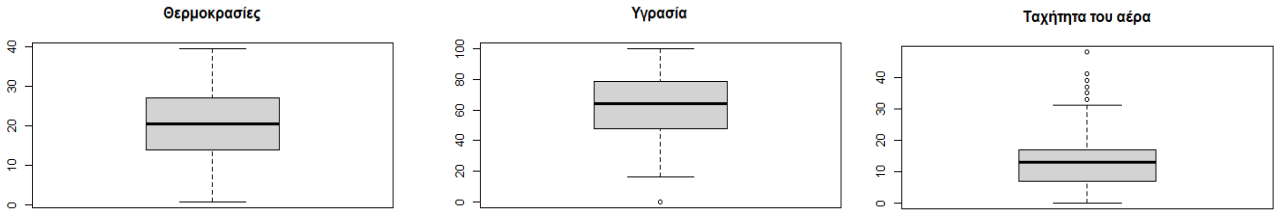
```
> summary(databike)
       X             instant          dteday           season       yr          mnth          hr          holiday       weekday
 Min.   :    5   Min.   :    5   Min.   :2011-01-01   1:376   0:753   7      :135   18     :  78   0:1458   0:217
 1st Qu.: 4282   1st Qu.: 4282   1st Qu.:2011-07-01   2:359   1:747   3      :134   4      :  73   1:  42   1:226
 Median : 8607   Median : 8607   Median :2011-12-30   3:406           12     :134   20     :  73            2:194
 Mean   : 8663   Mean   : 8663   Mean   :2012-01-01   4:359           5      :130   1      :  72            3:215
 3rd Qu.:13117   3rd Qu.:13117   3rd Qu.:2012-07-05                   9      :128   12     :  72            4:212
 Max.   :17378   Max.   :17378   Max.   :2012-12-31                   10     :128   8      :  71            5:212
                                                                      (other):711   (other):1061            6:224
 workingday    weathersit        temp            atemp              hum          windspeed          casual
 0: 483     Good    :974   Min.   : 0.82   Min.   : 1.515   Min.   :  0.00   Min.   : 0.000   Min.   :  0.00
 1:1017     Medium  :413   1st Qu.:13.94   1st Qu.:16.665   1st Qu.: 48.00   1st Qu.: 7.002   1st Qu.:  4.00
            Bad     :112   Median :20.50   Median :24.240   Median : 64.00   Median :12.998   Median : 16.00
            Very Bad:  1   Mean   :20.34   Mean   :23.755   Mean   : 63.33   Mean   :12.740   Mean   : 34.67
                           3rd Qu.:27.06   3rd Qu.:31.060   3rd Qu.: 79.00   3rd Qu.:16.998   3rd Qu.: 46.00
                           Max.   :39.36   Max.   :49.240   Max.   :100.00   Max.   :47.999   Max.   :350.00

  registered        cnt
 Min.   :  0.0   Min.   :  1.0
 1st Qu.: 35.0   1st Qu.: 41.0
 Median :115.5   Median :143.0
 Mean   :153.6   Mean   :188.3
 3rd Qu.:219.2   3rd Qu.:276.0
 Max.   :886.0   Max.   :977.0
```

Κάνοντας summary στο σετ δεδομένων μου, παρατηρώ ότι έχω 376 ενοικιάσεις που πραγματοποιήθηκαν τον Χειμώνα, 359 που πραγματοποιήθηκαν την Άνοιξη, 406 που πραγματοποιήθηκαν το Καλοκαίρι και 359 που πραγματοποιήθηκαν το Φθινόπωρο. Από αυτές τις ενοικιάσεις 753 πραγματοποιήθηκαν το 2011 και 747 το 2012. Μόλις 42 από τις 1500 αφορούν ημέρες διακοπών, ενώ 1017 από τις 1500 αφορούν εργάσιμες ημέρες. Οι περισσότερες ενοικιάσεις πραγματοποιήθηκαν σε ημέρες που επικρατούσαν καλές καιρικές συνθήκες (974 από 1500). Η μέση θερμοκρασία ήταν 20 βαθμούς ενώ η ελάχιστη ήταν 0.82 βαθμούς Κελσίου και η μέγιστη 39.36 βαθμούς. Η μέση υγρασία ήταν στους 63.33 βαθμούς ενώ η ελάχιστη στους 0 και η μέγιστη στους 100. Η μέση ταχύτητα του αέρα ήταν 12.74 ενώ η ελάχιστη ήταν 0 και η μέγιστη ήταν 48.



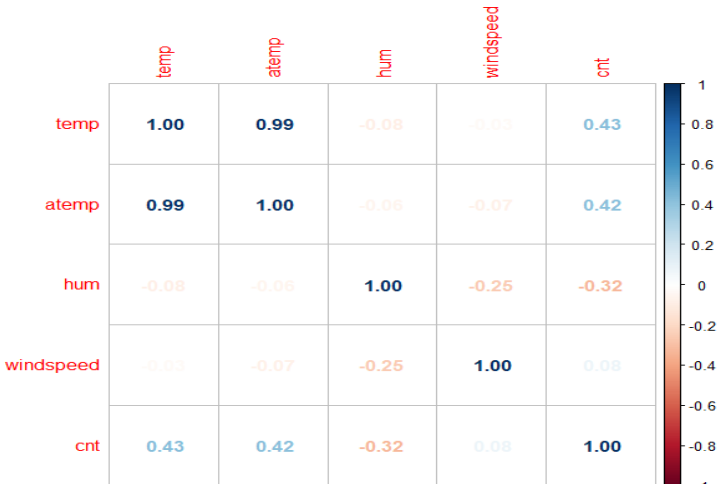
Παρατήρησα ότι οι αριθμητικές μεταβλητές δεν προέρχονται από κανονική κατανομή.



Διάγραμμα 0

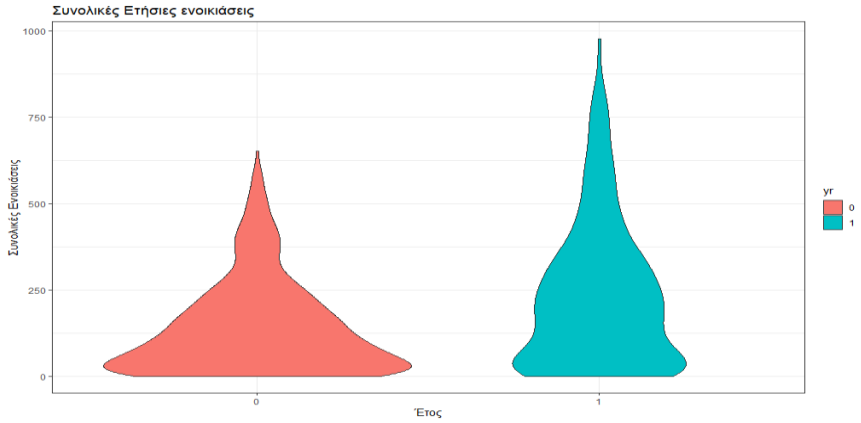
Από το Διάγραμμα 0, για τις ποσοτικές μεταβλητές, παρατήρησα ότι η μεταβλητή hum περιέχει μία ακραία τιμή. Πιο συγκεκριμένα, το hum παίρνει τιμή 0 σε μία παρατήρηση, κάτι που δεν είναι εφικτό να συμβαίνει. Και η μεταβλητή windspeed περιέχει 6 ακραίες τιμές που κατά πάσα πιθανότητα προέρχονται από ημέρες που δεν επικρατούσαν καλές καιρικές συνθήκες και είχε αρκετό αέρα.

Ανάλυση σχέσεων ανά δύο



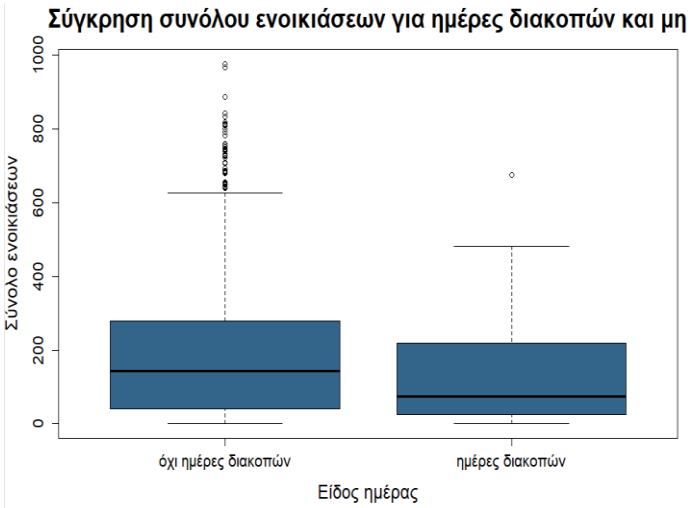
Διάγραμμα 1

Από το Διάγραμμα 1, κάνοντας δηλαδή corrplot για τις αριθμητικές μεταβλητές, παρατηρώ ότι το cnt έχει θετική γραμμική σχέση με το temp και το atemp, ενώ έχει αρνητική γραμμική σχέση με το hum. (για περισσότερη πληροφορία για τις σχέσεις των αριθμιτικών μεταβλητών δείτε στο Παράρτημα τα διαγράμματα 17 και 18 ενώ για περισσότερη πληροφορία σχετικά με τις κατηγορικές μεταβλητές και το cnt δείτε τα διαγράμματα 19 εως 26).

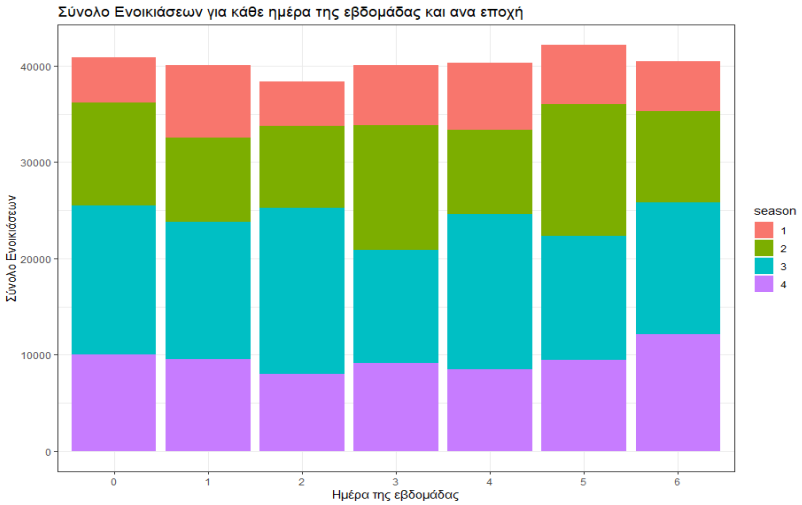


Διάγραμμα 2

Από το Διάγραμμα 2 φαίνεται να υπάρχει αυξητική τάση στις αυτοματοποιημένες ενοικιάσεις ποδηλάτων καθώς παρατηρούμε ότι το έτος 2012 οι ενοικιάσεις ποδηλάτων αυξήθηκαν αρκετά σε σχέση με το 2011.

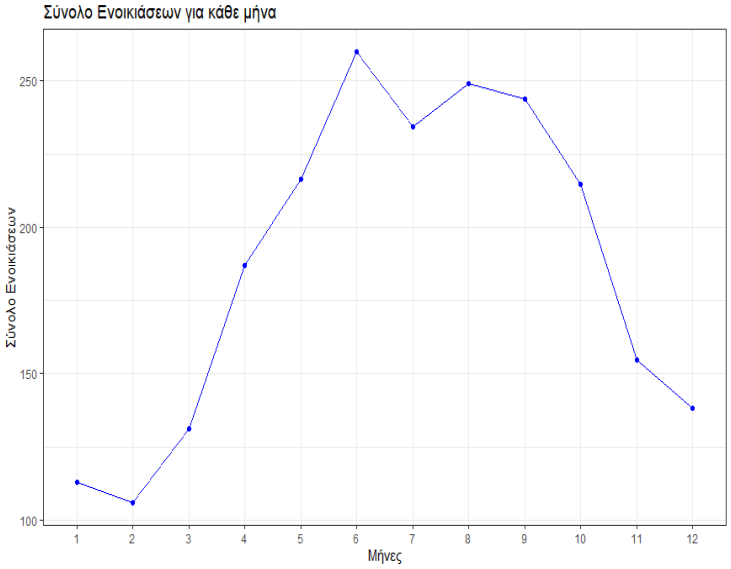


Διάγραμμα 9

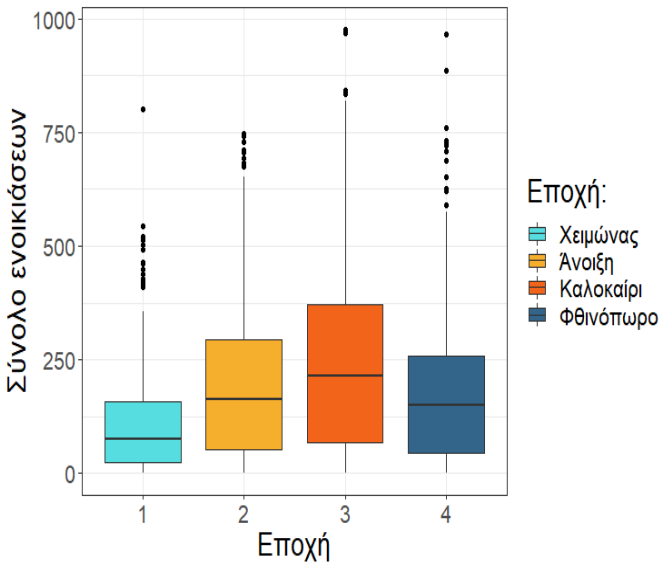


Διάγραμμα 7

Από το Διάγραμμα 7 και το Διάγραμμα 9 παρατηρώ ότι οι πολίτες χρησιμοποιούν σε καθημερινή βάση το ποδήλατο. Πιο συγκεκριμένα, στο Διάγραμμα 7 διαπιστώνω ότι οι πολίτες χρησιμοποιούν σε καθημερινή βάση το ποδήλατο τόσο για μετακινήσεις εντός πόλης, όσο και για βόλτα το Σαββατοκύριακο. Από το Διάγραμμα 9 συμπεραίνω πως παρ' ότι οι συνολικές ενοικιάσεις ανά ώρα στις ημέρες διακοπών και στις εργάσιμες ημέρες δεν διαφέρουν πολύ, υπάρχουν αρκετές ακραίες τιμές. Αυτές ,μάλλον, οφείλονται στο γεγονός ότι πολλοί άνθρωποι χρησιμοποιούν το ποδήλατο ως μέσω μετακίνησης για την δουλειά ειδικά κάποιες ημέρες οπού οι καιρικές συνθήκες είναι ευνοϊκές.

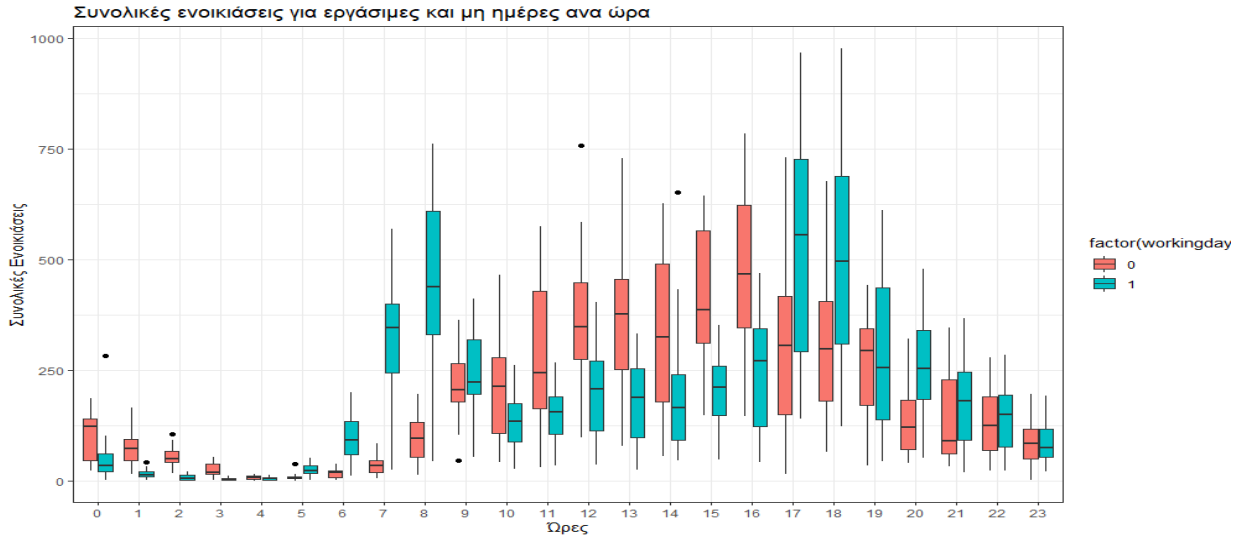


Διάγραμμα 8



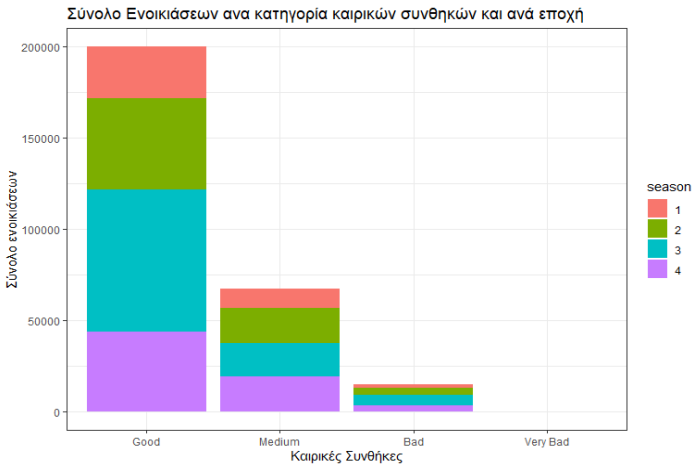
Διάγραμμα 12

Από το Διάγραμμα 12 παρατηρώ ότι την Άνοιξη και το Καλοκαίρι αυξάνονται οι ενοικιάσεις ποδηλάτων καθώς οι καιρικές συνθήκες είναι αρκετά πιο ευνοϊκές. Ορισμένες ακραίες τιμές που φαίνονται να υπάρχουν στο σύνολο ενοικιάσεων σε κάθε εποχή μπορεί να οφείλονται σε πολύ καλές ημέρες για την εκάστοτε εποχή ή μη εργάσιμες ημέρες σε συνδυασμό με μία σχετικά καλή ημέρα. Από το Διάγραμμα 8 παρατηρώ ότι τους μήνες Ιούνιο, Ιούλιο και Αύγουστο υπάρχει υψηλός αριθμός ενοικιάσεων, ενώ μετά τον Μάρτιο έως και τον Ιούνιο υπάρχει ραγδαία αύξηση στο σύνολο των ενοικιάσεων που φυσικά όπως προείπα οφείλεται στην βελτίωση των καιρικών συνθηκών και την αύξηση της θερμοκρασίας.

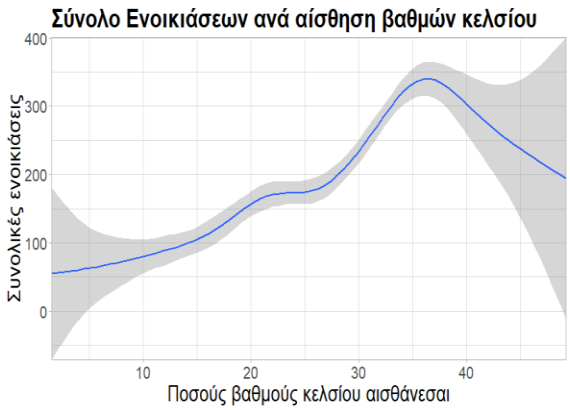


Διάγραμμα 3

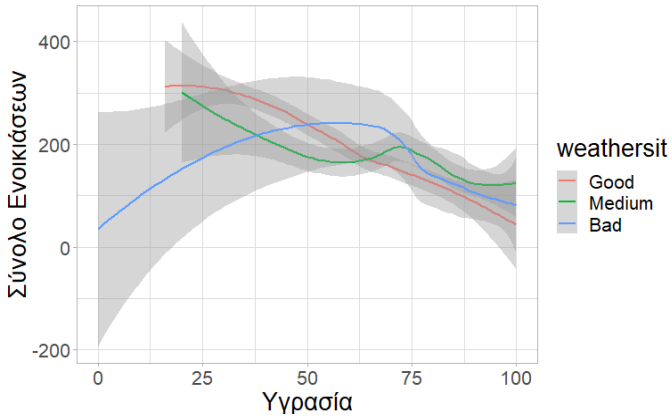
Από το Διάγραμμα 3 κατανοώ ότι τις εργάσιμες ημέρες παρατηρείται αύξηση στο σύνολο των ενοικιάσεων στις ώρες 7 και 8 το πρωί, επειδή οι πολίτες νοικιάζουν ποδήλατα για να πάνε στην εργασία τους. Επίσης, στις 5 και 6 το απόγευμα όπου σχολάνε και νοικιάζουν ποδήλατα για να επιστρέψουν στο σπίτι τους. Τα Σαββατοκύριακα οι ώρες που παρατηρείται αύξηση στο σύνολο των ενοικιάσεων είναι μεταξύ 11 το πρωί και 4 το μεσημέρι όπου χρησιμοποιούν το ποδήλατο για βόλτα.



Διάγραμμα 10

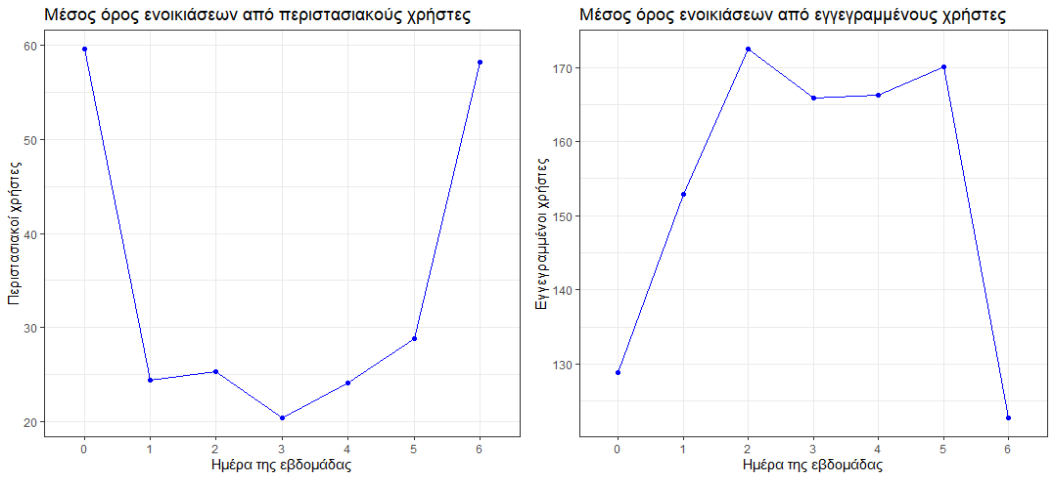


Διάγραμμα 13



Διάγραμμα 14

Από το Διάγραμμα 10, Διάγραμμα 13 και Διάγραμμα 14 παρατηρώ ότι οι συνολικές ενοικιάσεις είναι πολύ περισσότερες όταν οι καιρικές συνθήκες είναι ευνοϊκές και στις 4 εποχές και ότι οι πολίτες επιλέγουν να νοικιάζουν ποδήλατα για τις μετακινήσεις τους όσο η υγρασία και η θερμοκρασία δεν κυμαίνεται σε πολύ χαμηλά ή πολύ υψηλά επίπεδα.



Διάγραμμα 15

Από το Διάγραμμα 15 παρατηρώ ότι οι περιστασιακοί χρήστες ενοικιάζουν ποδήλατα για βόλτα κατά την διάρκεια του Σαββατοκύριακου, ενώ οι εγγεγραμμένοι χρήστες ενοικιάζουν ποδήλατα τις καθημερινές.

Μπορώ να συμπεράνω ότι μάλλον οι εγγεγραμμένοι χρήστες χρησιμοποιούν το ποδήλατο και ως μέσο μετακινήσεις για την δουλεία τους.

Μοντέλα Πρόβλεψης

Βήματα για την δημιουργία ενός μοντέλου πρόβλεψης

Για να μπορέσω να αναγνωρίσω το καλύτερο μοντέλο για την πρόβλεψη των συνολικών ενοικιάσεων για κάθε ώρα θα πρέπει να ξεκινήσω από ένα ολοκληρωμένο μοντέλο (fullmodel) το οποίο θα περιέχει την εξαρτημένη μεταβλητή Y = cnt και τις ανεξάρτητες μεταβλητές Xi = οι υπόλοιπες μεταβλητές εκτός από τις μεταβλητές που ήδη έχω αφαιρέσει. Επειδή το cnt είναι η εξαρτημένη μεταβλητή στο μοντέλο μου, εάν δεν αφαιρούσα τις μεταβλητές casual και registered θα καταλήγαμε σε ένα μοντέλο που θα περιέγραφε το προφανές με R-squared ίσο με 1. Αυτό συμβαίνει επειδή το cnt έχει υψηλή θετική γραμμική σχέση με το registered και το casual.

```
#remove the variables that are not usefull for the model
databikecleaned <- databike[,-c(1,2,3,16,17)]
```

```
> summary(fullmodel)

Call:
lm(formula = cnt ~ ., data = databikecleaned)
```

Το full γραμμικό μοντέλο στο νέο σετ δεδομένων

Επιλογή μεταβλητών με την τεχνική Lasso

Είναι πολύ σημαντικό το μοντέλο μου να περιέχει τις πιο σημαντικές μεταβλητές για την πρόβλεψη του Y. Για να το κάνουμε αυτό αρχικά χρησιμοποιούμε την τεχνική Lasso για να πετάξουμε από το μοντέλο τις στατιστικά μη σημαντικές μεταβλητές. Επιλέγω με βάση το lambda.1se = 2.74878 καθώς σε εκείνο το σημείο έχουν αφαιρεθεί αρκετές μεταβλητές από το μοντέλο. Ταυτόχρονα, το standard error είναι αρκετά μικρό (μικρότερο του ενός standard error). (για περισσότερες πληροφορίες σχετικά με την τεχνική Lasso), δείτε το Παράρτημα σελ 17). Το μοντέλο μου μετά την Lasso δεν περιέχει την μεταβλητή atemp και την μεταβλητή weekday.

```
> summary(model)

Call:
lm(formula = cnt ~ temp + hum + windspeed + season + yr + mnth +
    hr + holiday + workingday + weathersit, data = centerbikesdata)
```

Το μοντέλο μετά την τεχνική Lasso για lambda.1se = 2.74878

Επιλογή τελικών μεταβλητών μετά την Lasso με την τεχνική Stepwise

Στο μοντέλο που έχω μετά την Lasso θα χρησιμοποιήσω την τεχνική Stepwise με κατεύθυνση “both”, προκειμένου να παραμείνουν στο μοντέλο μου οι απολύτως απαραίτητες μεταβλητές X (για περισσότερες πληροφορίες σχετικά με την τεχνική Stepwise, δείτε το Παράρτημα σελ 17). Η μέθοδος αυτή προσθέτει και αφαιρεί κάθε φορά μεταβλητές βάση του κριτηρίου AIC μέχρι να καταλήξει στις απολύτως απαραίτητες μεταβλητές που πρέπει να παραμείνουν στο μοντέλο. Έτσι αφαιρώ από το μοντέλο μου την μεταβλητή “holiday” και το μοντέλο μας είναι το εξής:

Call: lm(formula = cnt ~ temp + hum + windspeed + season + yr + mnth + hr + workingday + weathersit, data = centerbikesdata)

Είναι πολύ σημαντικό μετά την τεχνική Stepwise να ελέγξω για πολυσυγγραμικότητα μεταξύ των μεταβλητών του μοντέλου. Η πολυσυγγραμικότητα εξετάζει την γραμμική σχέση κάθε μεταβλητής στο μοντέλο με τις υπόλοιπες μεταβλητές. Ο τρόπος για να αναγνωρίσουμε ποιες μεταβλητές πρέπει να αφαιρεθούν μετά τον έλεγχο είναι μέσω των δεικτών "VIF" = "Variance Inflation Factors". Εάν η τιμή του VIF για τις αριθμητικές μεταβλητές ή τις κατηγορικές μεταβλητές που αποτελούνται από το πολύ 2 κατηγορίες είναι μεγαλύτερη του 10, δημιουργούν πρόβλημα λόγω πολυσυγγραμικότητας και πρέπει να αφαιρεθούν από το μοντέλο. Ομοίως εάν η τιμή του VIF για τις κατηγορικές μεταβλητές που αποτελούνται από παραπάνω από 2 κατηγορίες είναι μεγαλύτερη του 3,16 επίσης δημιουργούν πρόβλημα λόγω πολυσυγγραμικότητας και πρέπει να αφαιρεθούν από το μοντέλο. Ελέγχοντας για πολυσυγγραμικότητα στο μοντέλο μου, παρατηρώ ότι η τιμή του VIF για δύο μεταβλητές είναι μεγαλύτερη από τα όρια που ανέφερα. Επιλέγω να αφαιρέσω την μεταβλητή με το μεγαλύτερο VIF (την μεταβλητή "mnth" με VIF = 387.30) από τις δύο και ξαναελέγχω για πολυσυγγραμικότητα. Παρατηρώ ότι η τιμή του VIF για την μεταβλητή "season" είναι λίγο μεγαλύτερη από τα όρια που ανέφερα (VIF = 3.437689). Αφαιρώ την μεταβλητή και ξαναελέγχω για πολυσυγγραμικότητα. Όλα τα VIF είναι μικρότερα του 10 και 3,16 για τις αντίστοιχες μεταβλητές. Άρα το μοντέλο μου είναι:

newmodel2 <- lm(formula = cnt ~ temp + hum + windspeed + yr + hr + workingday + weathersit, data = centerbikesdata)

Έλεγχο υποθέσεων στο τελικό μοντέλο newmodel2

Αφού έχω κάνει την τεχνική Lasso, την τεχνική Stepwise και έχω ελέγξει για πολυσυγγραμικότητα, έχω καταλήξει στο μοντέλο newmodel2, όπου πρέπει να ελέγξω εάν ικανοποιούνται οι παρακάτω προϋποθέσεις:

1. Κανονικότητα των σφαλμάτων : Τα σφάλματα πρέπει να είναι κανονικά
2. Ομοσκεδαστικότητα των σφαλμάτων : Τα σφάλματα πρέπει να έχουν ίδια διακύμανση σε κάθε επίπεδο των x
3. Ανεξαρτησία των σφαλμάτων : Να υπάρχει τυχαιότητα μεταξύ των σφαλμάτων
4. Γραμμικότητα των σφαλμάτων : Γραμμική σχέση μεταξύ των x και της εξαρτημένης μεταβλητής y

1. Για να ελέγξω την Κανονικότητα των σφαλμάτων κάνω Shapiro-Wilk και Lilliefors (Kolmogorov-Smirnov) μαζί με qqnorm. Και τα δύο τεστ υποθέσεων απορρίπτουν ότι τα σφάλματα είναι κανονικά.

Shapiro-Wilk: p-value < 2.2e-16 < 0.05, Lilliefors (Kolmogorov-Smirnov): p-value < 2.2e-16< 0.05

2. Για να ελέγξω την Ομοσκεδαστικότητα των σφαλμάτων κάνω ncvTest: p = < 2.22e-16.

3. Για να ελέγξω την Ανεξαρτησία των σφαλμάτων κάνω Runs Test και DurbinWatson Test.

Και τα δύο τεστ επιβεβαιώνουν ότι υπάρχει τυχαιότητα μεταξύ των σφαλμάτων.

Runs Test: p-value = 0.8768> 0.05 durbinWatsonTest: p-value = 0.52> 0.05

4. Για να ελέγξω την Γραμμικότητα των σφαλμάτων κάνω Tukey test όπου απορρίπτει την γραμμικότητα.

Tukey test: Pr(>|Test stat|) < 2e-16 ***< 0.05

Άρα ισχύει μόνο μία από τις τέσσερις προϋποθέσεις. Για να λύσω το πρόβλημα προσπαθώ να μετασχηματίσω το μοντέλο μου. Αρχικά εισάγω λογάριθμο στην εξαρτημένη μεταβλητή "cnt" (Y) και ξαναλέγχω τις προϋποθέσεις:

1. Shapiro-Wilk: p-value < 1.947e-12< 0.05, Lilliefors (Kolmogorov-Smirnov): p-value = 2.131e-12< 0.05
2. ncvTest: p = < 2.22e-16< 0.05
3. Runs Test: p-value = 0.7962> 0.05 durbinWatsonTest: p-value = 0.672> 0.05
4. Tukey test: Pr(>|Test stat) = 0.678> 0.05

Άρα ισχύουν δύο από τις τέσσερις προϋποθέσεις.

Έπειτα προσθέσω weight και ξαναλέγχω τις προϋποθέσεις:

1. Shapiro-Wilk: p-value < 9.003e-10< 0.05, Lilliefors (Kolmogorov-Smirnov): p-value = 4.073e-12< 0.05
2. ncvTest: p = 0.13731 > 0.05
3. Runs Test: p-value =0. 8768 > 0.05 durbinWatsonTest: p-value = 0.682 > 0.05
4. Tukey test: Pr(>|Test stat) = 0.72> 0.05

Άρα ισχύουν τρείς από τις τέσσερις προϋποθέσεις.

Αυτό είναι και το τελικό μου μοντέλο καθώς όταν πρόσθεσα πολυώνυμα και αφαίρεσα μεταβλητές δεν κατάφερα να ικανοποιήσω την κανονικότητα μάλλον επειδή δεν έχουμε στην κατοχή μας περισσότερα δεδομένα.

Ερμηνεία του τελικού μοντέλου το logmodel1

```
Call:
lm(formula = log(cnt) ~ temp + hum + windspeed + yr + hr + workingday +
    weathersit, data = databikecleaned, weights = w)

Weighted Residuals:
    Min      1Q  Median      3Q     Max
-5.1029 -0.7435  0.0885  0.8346  3.7130

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         2.8332166  0.1371520  20.657  < 2e-16 ***
temp                0.0418182  0.0019606  21.329  < 2e-16 ***
hum                 0.0006686  0.0010181   0.657 0.511465
windspeed          -0.0052224  0.0019294  -2.707 0.006872 **
yr1                 0.4737754  0.0298975  15.847  < 2e-16 ***
hr1                -0.8031021  0.1399372  -5.739 1.15e-08 ***
hr2                -1.3264472  0.1571483  -8.441  < 2e-16 ***
hr3                -2.1577452  0.1636458 -13.185  < 2e-16 ***
hr4                -2.3045041  0.1570174 -14.677  < 2e-16 ***
hr5                -1.0837902  0.1514802  -7.155 1.32e-12 ***
hr6                 0.0438319  0.1373484   0.319 0.749674
hr7                 1.1245094  0.1272937   8.834  < 2e-16 ***
hr8                 1.6530756  0.1182087  13.984  < 2e-16 ***
hr9                 1.4446661  0.1207580  11.963  < 2e-16 ***
hr10                1.0576081  0.1263701   8.369  < 2e-16 ***
hr11                1.2231014  0.1250951   9.777  < 2e-16 ***
hr12                1.4354003  0.1210707  11.856  < 2e-16 ***
hr13                1.4030131  0.1273807  11.014  < 2e-16 ***
hr14                1.4699958  0.1270418  11.571  < 2e-16 ***
hr15                1.2162921  0.1280556   9.498  < 2e-16 ***
hr16                1.6638971  0.1239283  13.426  < 2e-16 ***
hr17                1.9820010  0.1217918  16.274  < 2e-16 ***
hr18                1.9115526  0.1161306  16.460  < 2e-16 ***
hr19                1.5557926  0.1224968  12.701  < 2e-16 ***
hr20                1.3590293  0.1211074  11.222  < 2e-16 ***
hr21                1.0822605  0.1250985   8.651  < 2e-16 ***
hr22                0.8760911  0.1246504   7.028 3.19e-12 ***
hr23                0.3967023  0.1307723   3.034 0.002459 **
workingday1         0.1054122  0.0317767   3.317 0.000931 ***
weathersitMedium   -0.1364801  0.0375109  -3.638 0.000284 ***
weathersitBad      -0.6414102  0.0702275  -9.133  < 2e-16 ***
weathersitVery Bad -0.5931456  0.5563718  -1.066 0.286556
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.297 on 1468 degrees of freedom
Multiple R-squared:  0.7835,    Adjusted R-squared:  0.7789
F-statistic: 171.4 on 31 and 1468 DF,  p-value: < 2.2e-16
```

Από το summary για το μοντέλο μας, βλέπουμε ότι η intercept(βο) καθώς και οι μεταβλητές “temp”, “windspeed”,“yr1”,“hr1”, “hr2”, “hr3”, “hr4”, “hr5”, “hr7”, “hr8”, “hr9”, “hr10”, “hr11”, “hr12”, “hr13”, “hr14”, “hr15”, “hr16”, “hr17”, “hr18”, “hr19”, “hr20”, “hr21”, “hr22”, “hr23”, “workingday1”, “weathersitMedium” και “weathersitBad” είναι στατιστικά σημαντικές καθώς Pr(>|t|) < 0.05. Δηλαδή μόνο οι μεταβλητές “hum”, “hr6” και “weathersitVery Bad” δεν είναι στατιστικά σημαντικές. Επίσης βλέπουμε ότι το RMSE = 1.297 σε 1468 βαθμούς ελευθερίας. Αυτό σημαίνει ότι σε μία πρόβλεψη το μοντέλο θα αποκλίνει από την πραγματική τιμή του cnt(εξαρτημένης μεταβλητής) κατά exp(1.297) = 3.65%. Ακόμα, παρατηρώ ότι το R-squared είναι ίσο με 0.78. Αυτό σημαίνει ότι 78% τις μεταβλητότητας του συνόλου των ενοικιάσεων ποδηλάτου σε καθημερινή βάση μπορεί να εξηγηθεί από το μοντέλο.

Log(cnt) = 2.83 + 0.04 *temp + 0.00066 * hum – 0.0052 * windspeed + 0.47 *yr1 – 0.8 * hr1 – 1.32 * hr2 – 2.157 * hr3 – 2.3 * hr4– 1.08 * hr5+ 0.043* hr6 + 1.12 * hr7 + 1.65 * hr8 + 1.44 *hr9 + 1.057 * hr10 + 1.22 * hr11 + 1.43 * hr12 + 1.40 * hr13 + 1.47 * hr14 + 1.21* hr15 + 1.66 * hr16 + 1.99 * hr17 + 1.91 * hr18 + 1.55

* hr19 + 1.36 * hr20 + 1.08 * hr21 + 0.87 * hr22 + 0.39 * hr23 + 0.1 * workingday1 – 0.136 * weathersitMedium – 0.64 * weathersitBad – 0.59 * weathersit Very Bad + ε , ε ~ N(0, 1.297^2).

Το βο(intercept) = 2.83 είναι ο λογάριθμος της εξαρτημένης μεταβλητής "cnt" όταν όλες οι υπόλοιπες μεταβλητές είναι ίσες με μηδέν. Συνεπώς μπορώ να ερμηνεύσω το βο ως exp(2.83) = 16.94. Δηλαδή ότι έχουμε 16.94 συνολικές ενοικιάσεις ποδηλάτων όταν όλα τα υπόλοιπα χαρακτηρίστηκα είναι ίσα με το 0. Το β1 = 0.04. Άρα μία μονάδα αύξησης του "temp" αυξάνει τις συνολικές ενοικιάσεις ποδηλάτων κατά exp(0.04) = 1.04 = 4 % όταν όλες οι άλλες μεταβλητές είναι ίσες με 0. Το β2 = 0.00066 και συνεπώς μία μονάδα αύξησης του "hum " αυξάνει τις συνολικές ενοικιάσεις ποδηλάτων κατά exp(0.00066) = 1.00066021785= 0.6 % όταν όλες οι άλλες μεταβλητές είναι ίσες με 0. Το β3 = 0.00066 άρα μία μονάδα αύξησης του " windspeed " αυξάνει τις συνολικές ενοικιάσεις ποδηλάτων κατά – exp(0.0052) = – 1.00521354347= - 0.52 % όταν όλες οι άλλες μεταβλητές είναι ίσες με 0. Το β4 = 0.47. Άρα όταν οι υπόλοιπες μεταβλητές είναι ίσες με 0, στο έτος "yr1" = 2012 έχουμε αύξηση των συνολικών ενοικιάσεων κατά exp(0.47) = 1.5999 = 60% σε σχέση με το έτος "yr0" = 2011. Το β5 = – 0.8. Επομένως μπορώ να πω ότι στη 01:00 η ώρα = "hr1", οι συνολικές ενοικιάσεις ποδηλάτων είναι μειωμένες κατά - exp(0.8) = - 2.22 = 222% συγκριτικά με τις συνολικές ενοικιάσεις που πραγματοποιήθηκαν στο "hr0" = 00:00 η ώρα. Το β6 = – 1.32. Επομένως μπορώ να πω ότι στις 02:00 η ώρα = "hr2", οι συνολικές ενοικιάσεις ποδηλάτων είναι μειωμένες κατά - exp(1.32) = - 3.74= 374% συγκριτικά με τις συνολικές ενοικιάσεις που πραγματοποιήθηκαν στο "hr1" = 01:00 η ώρα. Το β7 = - 2.157. Επομένως μπορώ να πω ότι στις 03:00 η ώρα = "hr3" ,οι συνολικές ενοικιάσεις ποδηλάτων είναι μειωμένες κατά - exp(2.157) = - 8.64= 864% συγκριτικά με τις συνολικές ενοικιάσεις που πραγματοποιήθηκαν στο "hr2" = 02:00 η ώρα. Το β8 = – 2.3. Επομένως μπορώ να πω ότι στις 04:00 η ώρα = "hr4", οι συνολικές ενοικιάσεις ποδηλάτων είναι μειωμένες κατά - exp(2.3) = - 9.97= 997% συγκριτικά με τις συνολικές ενοικιάσεις που πραγματοποιήθηκαν στο "hr3" = 03:00 η ώρα. Το β9 = – 1.08. Επομένως μπορώ να πω ότι στις 05:00 η ώρα = "hr5", οι συνολικές ενοικιάσεις ποδηλάτων είναι μειωμένες κατά - exp(1.08) = - 2.94= 294% συγκριτικά με τις συνολικές ενοικιάσεις που πραγματοποιήθηκαν στο "hr4" = 04:00 η ώρα. Το β10 = 0.043. Επομένως μπορώ να πω ότι στις 06:00 η ώρα = "hr6" , οι συνολικές ενοικιάσεις ποδηλάτων είναι αυξημένες κατά exp(0.043) = 1.04= 4% συγκριτικά με τις συνολικές ενοικιάσεις που πραγματοποιήθηκαν στο "hr5" = 05:00 η ώρα.Το β11 = 1.12. Επομένως μπορώ να πω ότι στις 07:00 η ώρα = "hr7" , οι συνολικές ενοικιάσεις ποδηλάτων είναι αυξημένες κατά exp(1.12) = 3.06= 306% συγκριτικά με τις συνολικές ενοικιάσεις που πραγματοποιήθηκαν στο "hr6" = 06:00 η ώρα. Το β12 = 1.65. Επομένως μπορώ να πω ότι στις 08:00 η ώρα = "hr8" , οι συνολικές ενοικιάσεις ποδηλάτων είναι αυξημένες κατά exp(1.65) = 5.2= 520% συγκριτικά με τις συνολικές ενοικιάσεις που πραγματοποιήθηκαν στο "hr7" = 07:00 η ώρα.Το β13 = 1.65. Επομένως μπορώ να πω ότι στις 09:00 η ώρα = "hr9" =, οι συνολικές ενοικιάσεις ποδηλάτων είναι αυξημένες κατά exp(1.44) = 4.22= 422% συγκριτικά με τις συνολικές ενοικιάσεις που πραγματοποιήθηκαν στο "hr8" = 08:00 η ώρα. Το β14 = 1.057. Επομένως μπορώ να πω ότι στις 10:00 η ώρα = "hr10" , οι συνολικές ενοικιάσεις ποδηλάτων είναι αυξημένες κατά exp(1.057)= 2.87= 287% συγκριτικά με τις συνολικές ενοικιάσεις που πραγματοποιήθηκαν στο "hr9" = 09:00 η ώρα. Το β15 =1.22. Επομένως μπορώ να πω ότι στις 11:00 η ώρα = "hr11" οι συνολικές ενοικιάσεις ποδηλάτων είναι αυξημένες κατά exp(1.22)= 3.38= 338% συγκριτικά με τις συνολικές ενοικιάσεις που πραγματοποιήθηκαν στο "hr10" = 10:00 η ώρα. Το β16 =1.43. Επομένως μπορώ να πω ότι στις 12:00 η ώρα = "hr12", οι συνολικές ενοικιάσεις ποδηλάτων είναι αυξημένες κατά exp(1.43)= 4.17= 417% συγκριτικά με τις συνολικές ενοικιάσεις που πραγματοποιήθηκαν στο "hr11" = 11:00 η ώρα. Το β17 =1.40. Επομένως μπορώ να πω ότι στις 13:00 η ώρα "hr13" , οι συνολικές ενοικιάσεις ποδηλάτων είναι αυξημένες κατά exp(1.40)= 4.05= 405% συγκριτικά με τις συνολικές ενοικιάσεις που πραγματοποιήθηκαν στο "hr12" = 12:00 η ώρα. Το β18 = 1.47. Επομένως μπορώ να πω ότι στις 14:00 η ώρα "hr14" , οι συνολικές ενοικιάσεις ποδηλάτων είναι αυξημένες κατά exp(1.47)= 4.34= 434% συγκριτικά με τις συνολικές ενοικιάσεις που πραγματοποιήθηκαν στο "hr13" = 13:00 η ώρα. Το β19 = 1.21. Επομένως μπορώ να πω ότι στις 15:00 η ώρα "hr15" , οι συνολικές ενοικιάσεις ποδηλάτων είναι αυξημένες κατά exp(1.21) = 3.35 = 335% συγκριτικά με τις συνολικές ενοικιάσεις που πραγματοποιήθηκαν στο "hr14" = 14:00 η ώρα. Το β20 = 1.66. Επομένως μπορώ να πω ότι στις 16:00 η ώρα "hr16" , οι συνολικές ενοικιάσεις

ποδηλάτων είναι αυξημένες κατά exp(1.66) = 5.25 = 525% συγκριτικά με τις συνολικές ενοικιάσεις που πραγματοποιήθηκαν στο "hr15" = 15:00 η ώρα. Το β21 = 1.99. Επομένως μπορώ να πω ότι στις 17:00 η ώρα "hr17" , οι συνολικές ενοικιάσεις ποδηλάτων είναι αυξημένες κατά exp(1.99) = 7.31 = 731% συγκριτικά με τις συνολικές ενοικιάσεις που πραγματοποιήθηκαν στο "hr16" = 16:00 η ώρα. Το β22 = 1.91. Επομένως μπορώ να πω ότι στις 18:00 η ώρα "hr18" , οι συνολικές ενοικιάσεις ποδηλάτων είναι αυξημένες κατά exp(1.91) = 6.75 = 675% συγκριτικά με τις συνολικές ενοικιάσεις που πραγματοποιήθηκαν στο "hr17" = 17:00 η ώρα. Το β23 = 1.55. Επομένως μπορώ να πω ότι στις 19:00 η ώρα "hr19" , οι συνολικές ενοικιάσεις ποδηλάτων είναι αυξημένες κατά exp(1.55) = 4.71 = 471% συγκριτικά με τις συνολικές ενοικιάσεις που πραγματοποιήθηκαν στο "hr18" = 18:00 η ώρα. Το β24 = 1.36. Επομένως μπορώ να πω ότι στις 20:00 η ώρα "hr20" , οι συνολικές ενοικιάσεις ποδηλάτων είναι αυξημένες κατά exp(1.36) = 3.89 = 389% συγκριτικά με τις συνολικές ενοικιάσεις που πραγματοποιήθηκαν στο "hr19" = 19:00 η ώρα. Το β25 = 1.08. Επομένως μπορώ να πω ότι στις 21:00 η ώρα "hr21" , οι συνολικές ενοικιάσεις ποδηλάτων είναι αυξημένες κατά exp(1.08) = 2.94 = 294% συγκριτικά με τις συνολικές ενοικιάσεις που πραγματοποιήθηκαν στο "hr20" = 20:00 η ώρα. Το β26 = 0.87. Επομένως μπορώ να πω ότι στις 22:00 η ώρα "hr22" , οι συνολικές ενοικιάσεις ποδηλάτων είναι αυξημένες κατά exp(0.87)= 2.38 = 238% συγκριτικά με τις συνολικές ενοικιάσεις που πραγματοποιήθηκαν στο "hr21" = 21:00 η ώρα. Το β27 = 0.39. Επομένως μπορώ να πω ότι στις 23:00 η ώρα "hr23" , οι συνολικές ενοικιάσεις ποδηλάτων είναι αυξημένες κατά exp(0.39)= 1.47 = 147% συγκριτικά με τις συνολικές ενοικιάσεις που πραγματοποιήθηκαν στο "hr22" = 22:00 η ώρα. Το β28 = 0.1. Άρα όταν οι υπόλοιπες μεταβλητές είναι ίσες με 0, στις εργάσιμες ημέρες " workingday1" έχουμε αύξηση των συνολικών ενοικιάσεων κατά exp(0.1) = 1.10 = 10% σε σχέση με τις μη εργάσιμες ημέρες "workingday0". Το β29 = - 0.136. Άρα όταν οι υπόλοιπες μεταβλητές είναι ίσες με 0, στις ημέρες που επικρατούν μέτριες καιρικές συνθήκες "weathersitMedium" έχουμε μείωση των συνολικών ενοικιάσεων κατά - exp(0.136) = - 1.14 = 14% σε σχέση με τις ημέρες που επικρατούν καλές καιρικές συνθήκες "weathersitGood". Το β30 = – 0.64. Άρα όταν οι υπόλοιπες μεταβλητές είναι ίσες με 0, στις ημέρες που επικρατούν κακές καιρικές συνθήκες "weathersitBad" έχουμε μείωση των συνολικών ενοικιάσεων κατά - exp(0.64) = - 1.89 = 89% σε σχέση με τις ημέρες όπου επικρατούν μέτριες καιρικές συνθήκες "weathersitMedium". Το β31 = – 0.59. Άρα όταν οι υπόλοιπες μεταβλητές είναι ίσες με 0, στις ημέρες που επικρατούν πολύ κακές καιρικές συνθήκες "weathersit Very Bad" έχουμε μείωση των συνολικών ενοικιάσεων κατά - exp(0.59) = - 1.80 = 80% σε σχέση με τις ημέρες όπου επικρατούν κακές καιρικές συνθήκες "weathersitBad ".

Ικανότητα Πρόβλεψης

Θα προσπαθήσουμε να διαλέξουμε μεταξύ του fullmodel, του Nullmodel και του logmodel μοντέλου μου, αυτό που είναι καλύτερο για πρόβλεψη σε άλλα σετ δεδομένων. Πιο συγκεκριμένα, θα χρησιμοποιήσω το σετ δεδομένων test_data που περιέχει 500 παρατηρήσεις και θα προσπαθήσω να κάνω πρόβλεψη με τα 3 αυτά μοντέλα. Έπειτα θα συγκρίνω το mae = " Mean Absolute Error" για να δω πιο από τα 3 μοντέλα έχει μικρότερο mae και επομένως είναι καλύτερο για πρόβλεψη σε άλλα σετ δεδομένων. Το Mean Absolute Error(mae) είναι ο μέσος όρος της απόλυτης τιμής της διαφοράς των πραγματικών τιμών και των τιμών που πρόβλεψε το μοντέλο μας. Στην Εικόνα 1 εμφανίζονται τα mae των 3 μοντέλων για την πρόβλεψη στο σετ δεδομένων test_data.

```
> #Task 5
> nullmodel <- lm(formula = cnt ~ 1 , data = databikecleaned)
> fullmodel <- lm(formula = cnt ~ . , data = databikecleaned)
> logmodel <- lm(formula = log(cnt) ~ 1 + yr + hr + weathersit +
+                  temp + hum + I(hum^2) + I(temp^2) ,weight = w, data = databikecleaned)
> mae(test_data$cnt, predict(nullmodel))
[1] 145.5478
> mae(test_data$cnt, predict(fullmodel))
[1] 182.6984
> mae(test_data$cnt, predict(logmodel))
[1] 180.9328
```

Εικόνα 1 MAE των τριών μοντέλων

Από την Εικόνα 1 παρατηρώ ότι το μηδενικο μοντέλο nullmodel έχει το μικρότερο Mean Absolute Error και άρα είναι το καλύτερο μοντέλο από τα 3 για πρόβλεψη σε άλλα σετ δεδομένων καθώς όσο μικρότερο είναι το mae τόσο καλύτερη είναι η πρόβλεψη.

Επιπλέον Ανάλυση

Σε αυτή την ενότητα θα προσπαθήσω να περιγράψω μία τυπική ημέρα για κάθε μία από τις 4 εποχές του χρόνου σύμφωνα με το αρχικό σετ δεδομένων "databike". Αφού πρώτα δημιούργησα 4 υποσύνολα του αρχικού μου σετ δεδομένων(1 υποσύνολο για κάθε εποχή που περιέχει τις παρατηρήσεις από τις ενοικιάσεις που πραγματοποιήθηκαν σε αυτή την εποχή του χρόνου).

Περιγραφή μίας τυπικής ημέρας τον Χειμώνα

```
> s1 = subset(databike, season == 1)
> summary(s1)
       X             instant          dteday           season  yr           mnth          hr       holiday weekday
 Min.   :    5   Min.   :    5   Min.   :2011-01-01   1:376   0:192   1      :123   4      : 21   0:362   0:47
 1st Qu.: 1180   1st Qu.: 1180   1st Qu.:2011-02-21   2:  0   1:184   2      :112   18     : 21   1: 14   1:64
 Median : 8578   Median : 8578   Median :2011-12-29   3:  0           3      : 94   14     : 20           2:48
 Mean   : 6132   Mean   : 6132   Mean   :2011-09-16   4:  0           12     : 47   19     : 20           3:63
 3rd Qu.: 9723   3rd Qu.: 9723   3rd Qu.:2012-02-15                   4      :  0   20     : 20           4:54
 Max.   :17378   Max.   :17378   Max.   :2012-12-31                   5      :  0   12     : 19           5:50
                                                                      (Other):  0   (Other):255           6:50
 workingday    weathersit       temp           atemp             hum          windspeed          casual
 0:111      Good    :234   Min.   : 0.82   Min.   : 1.515   Min.   :  0.0   Min.   : 0.000   Min.   :  0.00
 1:265      Medium  :114   1st Qu.: 9.02   1st Qu.:11.365   1st Qu.: 44.0   1st Qu.: 7.002   1st Qu.:  1.00
            Bad     : 27   Median :11.48   Median :13.635   Median : 57.0   Median :15.001   Median :  5.00
            Very Bad:  1   Mean   :12.32   Mean   :14.996   Mean   : 58.4   Mean   :14.479   Mean   : 13.18
                           3rd Qu.:15.58   3rd Qu.:19.695   3rd Qu.: 70.0   3rd Qu.:20.000   3rd Qu.: 13.00
                           Max.   :29.52   Max.   :32.575   Max.   :100.0   Max.   :47.999   Max.   :176.00

  registered         cnt
 Min.   :  1.0   Min.   :  1.0
 1st Qu.: 23.0   1st Qu.: 24.5
 Median : 65.0   Median : 76.5
 Mean   : 96.8   Mean   :110.0
 3rd Qu.:139.2   3rd Qu.:157.5
 Max.   :681.0   Max.   :801.0
```

Εικόνα 2: Περίληψη του υποσυνόλου για τον Χειμώνα

Παρατηρώντας την Εικόνα 2, μπορώ να πω ότι μία τυπική ημέρα του Χειμώνα έχει 12.32 βαθμούς Κελσίου (παρατηρώντας ότι υπήρξε και ημέρα με 29.5 βαθμούς Κελσίου) , η αίσθηση της θερμοκρασίας είναι στους 14.99 βαθμούς Κελσίου, ενώ η ταχύτητα του αέρα ήταν 14.47. Ακόμα ο μέσος όρος των ενοικιάσεων είναι 110 από τις οποίες οι 13 πραγματοποιούνται από περιστασιακούς χρήστες ενώ οι 97 από εγγεγραμμένους χρήστες.

Περιγραφή μίας τυπικής ημέρας την Άνοιξη

```
> s2 = subset(databike, season == 2)
> summary(s2)
       X            instant          dteday           season     yr          mnth          hr          holiday     weekday
 Min.   : 1811   Min.   : 1811   Min.   :2011-03-21   1:  0   0:182   5      :130   22     : 24   0:353   0:49
 1st Qu.: 2904   1st Qu.: 2904   1st Qu.:2011-05-05   2:359   1:177   4      :109   8      : 23   1:  6   1:53
 Median : 3985   Median : 3985   Median :2011-06-19   3:  0           6      : 80   21     : 22           2:47
 Mean   : 7238   Mean   : 7238   Mean   :2011-11-02   4:  0           3      : 40   12     : 20           3:61
 3rd Qu.:11672   3rd Qu.:11672   3rd Qu.:2012-05-06                   1      :  0   18     : 18           4:41
 Max.   :12755   Max.   :12755   Max.   :2012-06-20                   2      :  0   23     : 18           5:56
                                                                      (Other):  0   (Other):234           6:52
  workingday    weathersit       temp           atemp             hum          windspeed          casual
 0:107      Good     :216   Min.   : 8.20   Min.   : 9.85   Min.   : 20.0   Min.   : 0.000   Min.   :  0.0
 1:252      Medium   :108   1st Qu.:18.86   1st Qu.:22.73   1st Qu.: 48.0   1st Qu.: 8.998   1st Qu.:  7.0
            Bad      : 35   Median :22.96   Median :25.76   Median : 65.0   Median :12.998   Median : 24.0
            Very Bad:  0   Mean   :22.30   Mean   :25.93   Mean   : 64.7   Mean   :13.596   Mean   : 41.4
                           3rd Qu.:26.24   3rd Qu.:31.06   3rd Qu.: 83.0   3rd Qu.:16.998   3rd Qu.: 55.0
                           Max.   :36.90   Max.   :40.91   Max.   :100.0   Max.   :39.001   Max.   :317.0

   registered        cnt
 Min.   :  0.0   Min.   :  1.0
 1st Qu.: 41.0   1st Qu.: 52.0
 Median :130.0   Median :162.0
 Mean   :161.8   Mean   :203.2
 3rd Qu.:220.5   3rd Qu.:295.5
 Max.   :675.0   Max.   :748.0
```

Εικόνα 3: Περίληψη του υποσυνόλου για την Άνοιξη

Παρατηρώντας την Εικόνα 3, μπορώ να πω ότι μία τυπική ημέρα την Άνοιξη έχει 22.30 βαθμούς Κελσίου(παρατηρώντας ότι υπήρξε ημέρα με 36.9 βαθμούς Κελσίου αλλά και ημέρα με 8.2 βαθμούς Κελσίου), η αίσθηση της θερμοκρασίας είναι στους 25.93 βαθμούς Κελσίου, ενώ η ταχύτητα του αέρα είναι 13.6. Ακόμα ο μέσος όρος των ενοικιάσεων είναι 203 από τις οποίες οι 41 πραγματοποιούνται από περιστασιακούς χρήστες ενώ οι 162 από εγγεγραμμένους χρήστες. Συνολικά παρατηρώ ότι οι ενοικιάσεις των ποδηλάτων είναι διπλάσιες από ότι τον Χειμώνα και αυτό οφείλεται στις καλύτερες καιρικές συνθήκες που επικρατούν την Άνοιξη.

Περιγραφή μίας τυπικής ημέρας το Καλοκαίρι

```
> s3 = subset(databike, season == 3)
> summary(s3)
       X            instant          dteday           season     yr          mnth          hr          holiday     weekday
 Min.   : 4012   Min.   : 4012   Min.   :2011-06-21   1:  0   0:196   7      :135   1      : 25   0:397   0:64
 1st Qu.: 5220   1st Qu.: 5220   1st Qu.:2011-08-10   2:  0   1:210   8      :125   18     : 25   1:  9   1:57
 Median :12832   Median :12832   Median :2012-06-23   3:406           9      :100   5      : 22           2:58
 Mean   : 9655   Mean   : 9655   Mean   :2012-02-11   4:  0           6      : 46   8      : 22           3:46
 3rd Qu.:13872   3rd Qu.:13872   3rd Qu.:2012-08-06                   1      :  0   9      : 21           4:67
 Max.   :15014   Max.   :15014   Max.   :2012-09-22                   2      :  0   4      : 20           5:55
                                                                      (Other):  0   (Other):271           6:59
  workingday    weathersit       temp           atemp             hum          windspeed          casual
 0:132      Good     :305   Min.   :15.58   Min.   :12.12   Min.   : 19.00   Min.   : 0.000   Min.   :  0.00
 1:274      Medium   : 80   1st Qu.:26.24   1st Qu.:30.30   1st Qu.: 49.25   1st Qu.: 7.002   1st Qu.: 11.00
            Bad      : 21   Median :28.70   Median :32.58   Median : 65.00   Median :11.001   Median : 35.00
            Very Bad:  0   Mean   :28.85   Mean   :32.66   Mean   : 62.65   Mean   :11.375   Mean   : 50.85
                           3rd Qu.:31.16   3rd Qu.:34.85   3rd Qu.: 78.00   3rd Qu.:16.998   3rd Qu.: 69.00
                           Max.   :39.36   Max.   :49.24   Max.   :100.00   Max.   :39.001   Max.   :350.00

   registered        cnt
 Min.   :  2.0   Min.   :  2.0
 1st Qu.: 54.0   1st Qu.: 69.0
 Median :161.0   Median :214.5
 Mean   :198.5   Mean   :249.3
 3rd Qu.:282.0   3rd Qu.:372.2
 Max.   :886.0   Max.   :977.0
```

Εικόνα 4: Περίληψη του υποσυνόλου για το Καλοκαίρι

Παρατηρώντας την Εικόνα 4, μπορώ να πω ότι μία τυπική ημέρα το Καλοκαίρι έχει 28.85 βαθμούς Κελσίου(παρατηρώντας ότι υπήρξε και ημέρα με 15.58 βαθμούς Κελσίου), η αίσθηση της θερμοκρασίας είναι στους 32.66 βαθμούς Κελσίου, ενώ η ταχύτητα του αέρα είναι 11.37. Ακόμα ο μέσος όρος των ενοικιάσεων είναι 249 από τις οποίες οι 51 πραγματοποιούνται από περιστασιακούς χρήστες ενώ οι 198 από εγγεγραμμένους χρήστες. Συνολικά παρατηρώ ότι οι ενοικιάσεις των ποδηλάτων είναι σχεδόν δυόμιση φορές περισσότερες απ' ότι τον Χειμώνα και ¼ περισσότερες απ' ότι την Άνοιξη.

Περιγραφή μίας τυπικής ημέρας το Φθινόπωρο

```
> s4 = subset(databike, season == 4)
> summary(s4)
       X             instant          dteday           season  yr           mnth          hr       holiday weekday
 Min.   : 6257   Min.   : 6257   Min.   :2011-09-23   1:  0   0:183   10     :128   6      : 20   0:346   0:57
 1st Qu.: 7268   1st Qu.: 7268   1st Qu.:2011-11-04   2:  0   1:176   11     :116   16     : 20   1: 13   1:52
 Median : 8314   Median : 8314   Median :2011-12-18   3:  0           12     : 87   7      : 18           2:41
 Mean   :11618   Mean   :11618   Mean   :2012-05-03   4:359           9      : 28   14     : 18           3:45
 3rd Qu.:16087   3rd Qu.:16087   3rd Qu.:2012-11-07                   1      :  0   20     : 18           4:50
 Max.   :17106   Max.   :17106   Max.   :2012-12-20                   2      :  0   4      : 17           5:51
                                                                     (Other):  0   (Other):248           6:63
 workingday    weathersit       temp           atemp             hum          windspeed          casual
 0:133      Good    :219   Min.   : 5.74   Min.   : 7.575   Min.   : 29.00   Min.   : 0.000   Min.   :  0.00
 1:226      Medium  :111   1st Qu.:13.94   1st Qu.:16.665   1st Qu.: 54.00   1st Qu.: 6.003   1st Qu.:  4.00
            Bad     : 29   Median :16.40   Median :20.455   Median : 69.00   Median :11.001   Median : 13.00
            Very Bad:  0   Mean   :17.18   Mean   :20.686   Mean   : 67.92   Mean   :11.605   Mean   : 32.14
                           3rd Qu.:21.32   3rd Qu.:25.000   3rd Qu.: 82.00   3rd Qu.:16.998   3rd Qu.: 38.50
                           Max.   :30.34   Max.   :33.335   Max.   :100.00   Max.   :35.001   Max.   :347.00

  registered        cnt
 Min.   :  1.0   Min.   :  1.0
 1st Qu.: 40.5   1st Qu.: 46.0
 Median :126.0   Median :151.0
 Mean   :154.3   Mean   :186.4
 3rd Qu.:222.5   3rd Qu.:258.5
 Max.   :860.0   Max.   :967.0
```

Εικόνα 5: Περίληψη του υποσυνόλου για το Φθινόπωρο

Παρατηρώντας την Εικόνα 5, μπορώ να πω ότι μία τυπική ημέρα το Φθινόπωρο έχει 17.18 βαθμούς Κελσίου(παρατηρώντας ότι υπήρξε ημέρα με 5.74 βαθμούς Κελσίου αλλά και ημέρα με 30.34 βαθμούς Κελσίου, η αίσθηση της θερμοκρασίας είναι στους 20.68 βαθμούς Κελσίου, ενώ η ταχύτητα του αέρα είναι 11.6. Ακόμα ο μέσος όρος των ενοικιάσεων είναι 186 από τις οποίες οι 13 πραγματοποιούνται από περιστασιακούς χρήστες ενώ οι 154 από εγγεγραμμένους χρήστες. Συνολικά παρατηρώ ότι οι ενοικιάσεις των ποδηλάτων είναι σχεδόν 1/2 φορές περισσότερες απ' ότι τον Χειμώνα και ¼ λιγότερες απ' ότι την Άνοιξη.

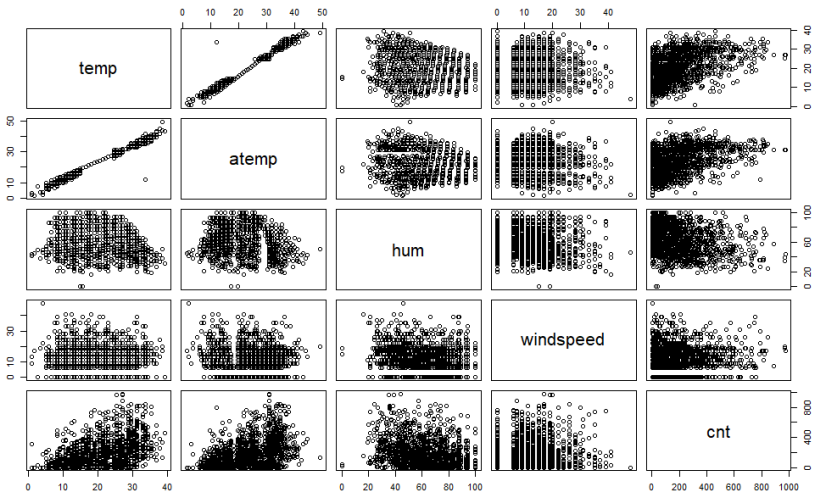
Παράρτημα

Lasso

Στη μέθοδο LASSO με την αύξηση της τιμής της παραμέτρου λ οι συντελεστές παλινδρόμησης συρρικνώνονται. Το πλεονέκτημα όμως της μεθόδου αυτής σε σχέση με την μέθοδο Ridge είναι ότι κάποιοι συντελεστές παλινδρόμησης μπορούν να γίνουν ακριβώς ίσοι με το μηδέν. Αυτό επιτυγχάνεται λόγω του περιορισμού ∑ |βk | ≤ t p k=1 στην ελαχιστοποίηση της παράστασης. Γεωμετρικά ο περιορισμός ∑ |βk | ≤ t p k=1 παριστάνει ένα πολύτοπο (polytope) με πολλές κορυφές. Οι κορυφές αυτές είναι τα σημεία όπου κάποιοι συντελεστές παίρνουν την τιμή 0. Άμεσο επακόλουθο των παραπάνω είναι ότι η ελαχιστοποίηση της παράστασης ∑ (yi − β0 − ∑ $\beta kxik$ p k=1) 2 n i=1 υπό τον περιορισμό ∑ |βk | ≤ t p k=1 , για κάποια τιμή του t, μπορεί να επιτυγχάνεται σε μία κορυφή του πολύτοπου, συνεπώς κάποιοι συντελεστές είναι ίσοι με το 0.

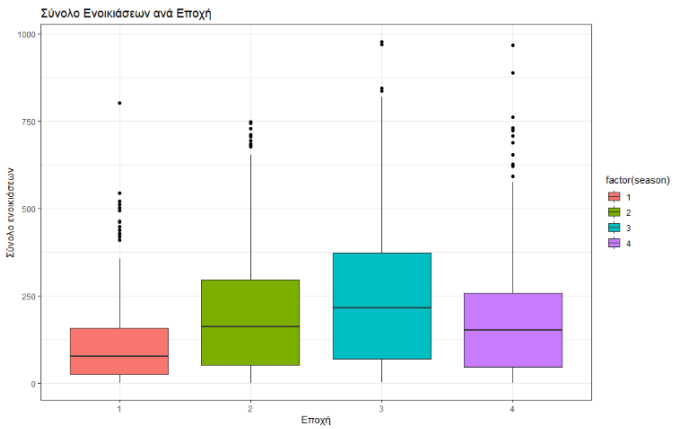
Stepwise Regression

Η μέθοδος της βηματικής παλινδρόμησης (stepwise regression) είναι μιά άλλη μέθοδος επιλογής ενός "καλού" υποσυνόλου ανεξαρτήτων μεταβλητών. Η μέθοδος αυτή είναι παρόμοια με την μέθοδο της προοδευτικής προσθήκης μεταβλητών. Η διαφορά των δύο μεθόδων έγκειται στο γεγονός ότι για κάθε διαδοχικό βήμα η υπόθεση H0 : βj = 0 ελέγχεται για όλες τις δυνατές ανεξάρτητες μεταβλητές ώστε να αποκλείονται εκείνες για τις οποίες οι τιμές της στατιστικής συνάρτησης |Tj| είναι μικρότερες από ένα προκαθορισμένο κρίσιμο επίπεδο. Η επόμενη μεταβλητή προστίθεται στο υποσύνολο με την ίδια διαδικασία χρησιμοποίησης του κριτηρίου του μεγίστου συντελεστή συσχέτισης όπως στη μέθοδο της προοδευτικής προσθήκης μεταβλητών. Αυτή η βηματική επιλογή συνεχίζεται μέχρις ότου φθάσουμε σε ένα υποσύνολο μεταβλητών για το οποίο καμμιά από τις μεταβλητές που περιέχει το υποσύνολο αυτό δεν έχουν τιμή για τη στατιστική συνάρτηση |Tj| μικρότερη από κάποια συγκεκριμένη κρίσιμη τιμή της μεταβλητής t και δεν υπάρχουν άλλες μεταβλητές που θα πρέπει να αξιολογηθούν για να περιληφθούν στο μοντέλο.

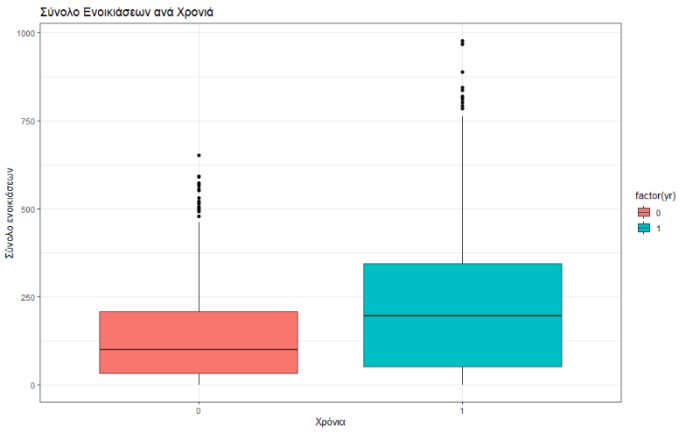


Διάγραμμα 17

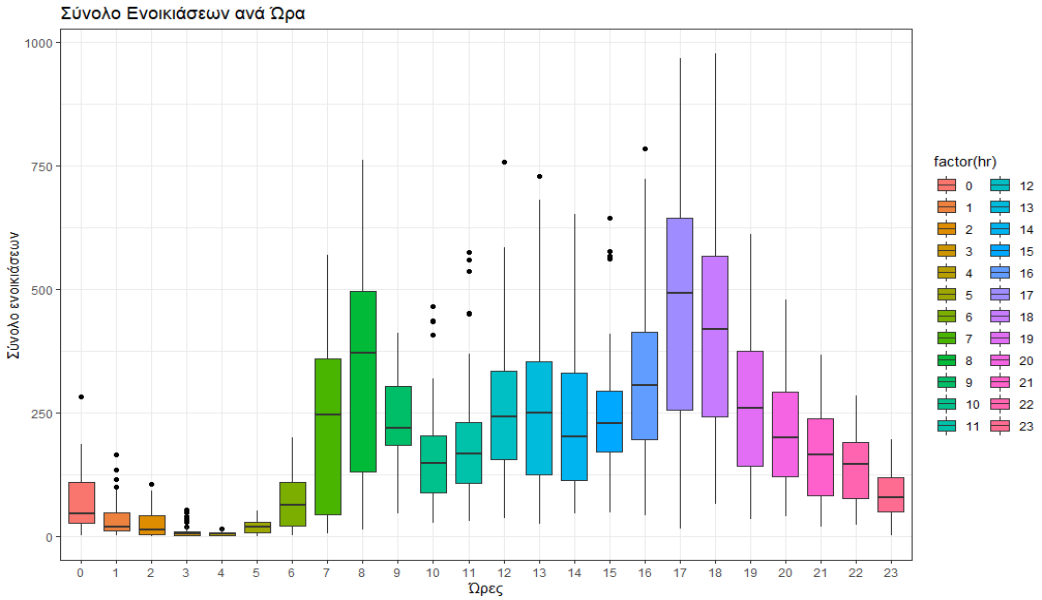
Διάγραμμα 18



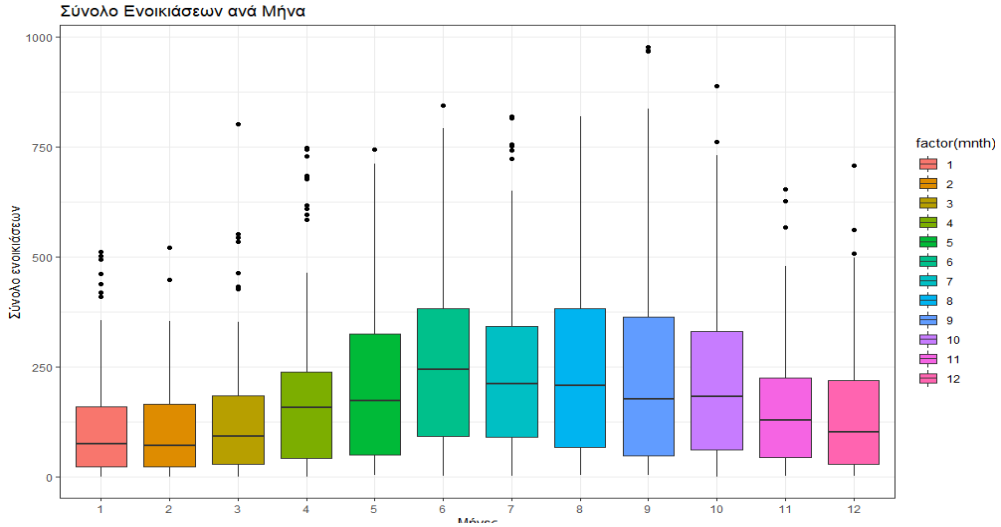
Διάγραμμα 19



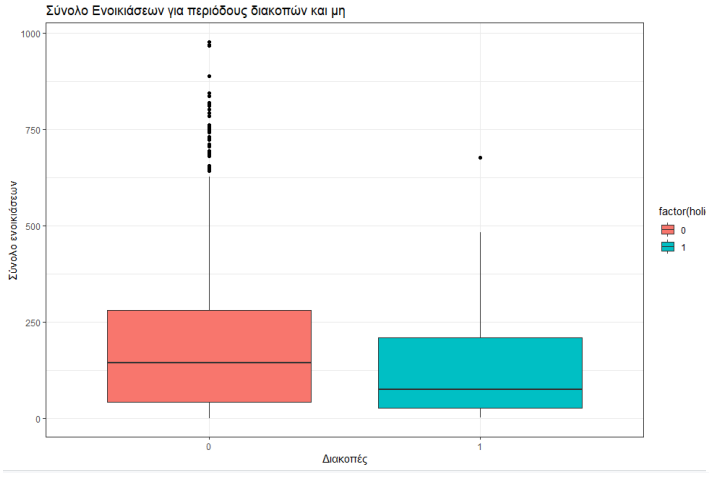
Διάγραμμα 20



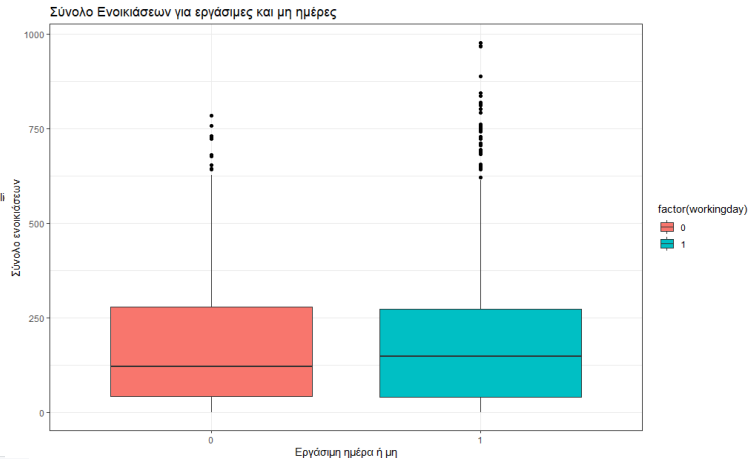
Διάγραμμα 21



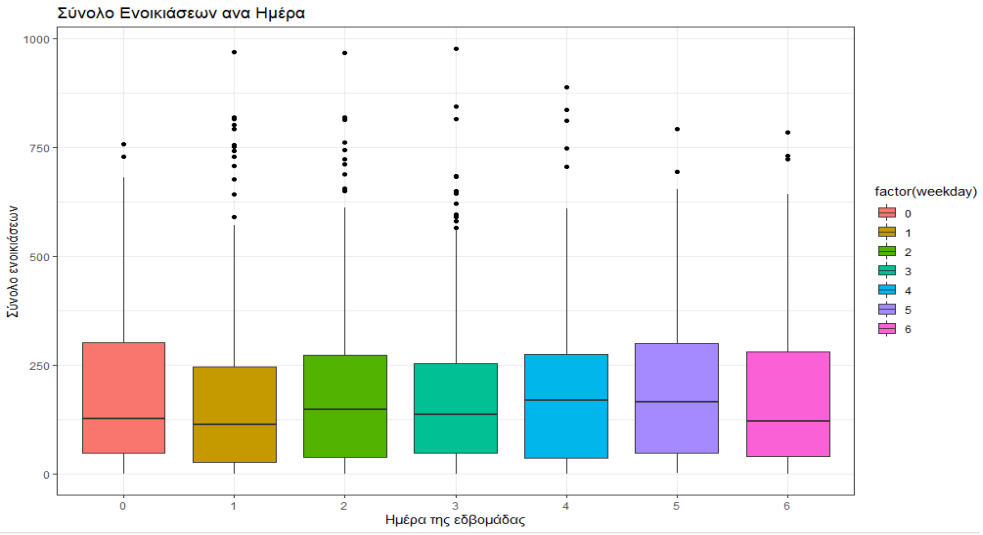
Διάγραμμα 22



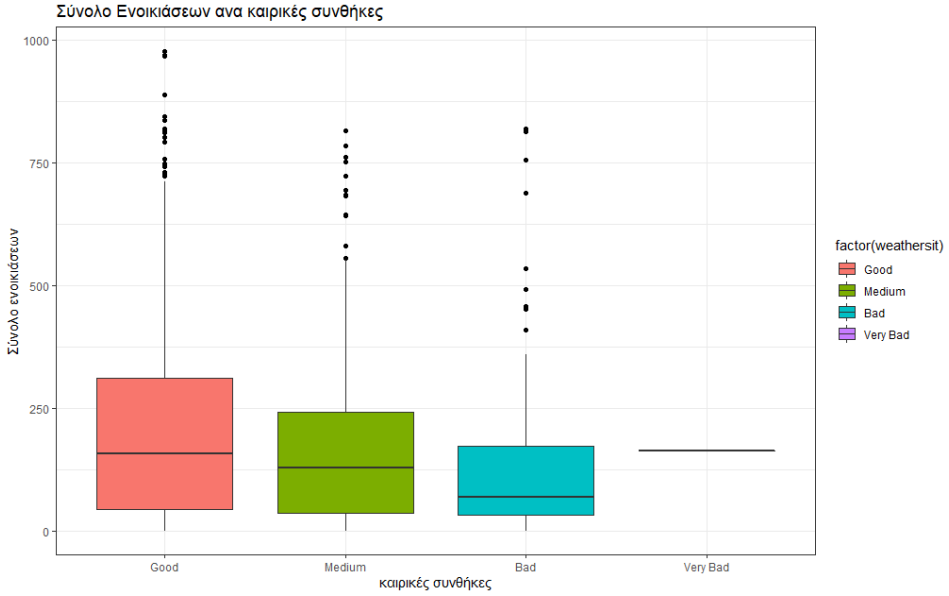
Διάγραμμα 23



Διάγραμμα 24



Διάγραμμα 25



Διάγραμμα 26