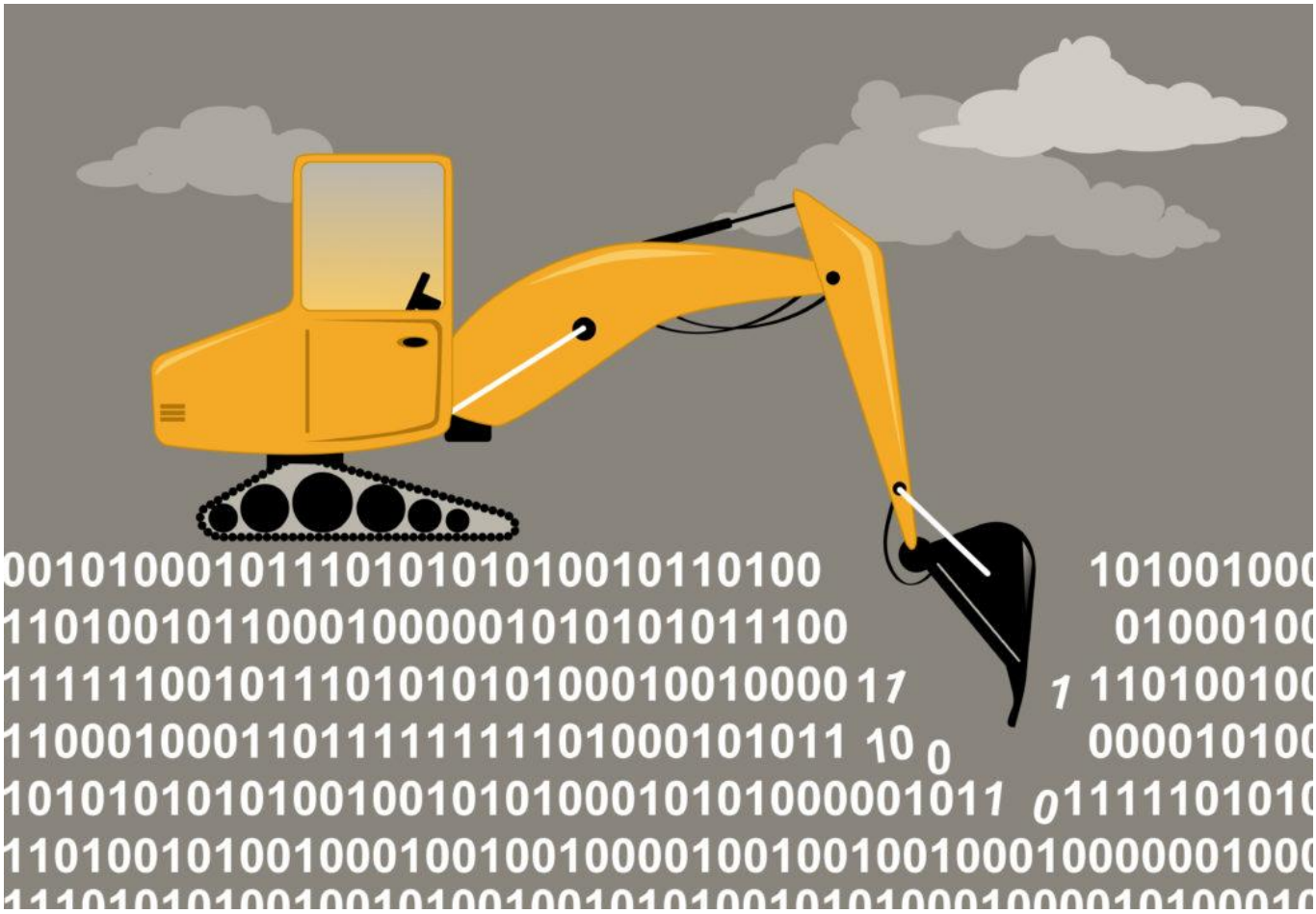
Μάθημα: «Αποθήκες Δεδομένων και Εξόρυξη Γνώσης (6ο εξ.)»

Ομάδα εργασίας: (Π15128, Σκάρος Γεώργιος)

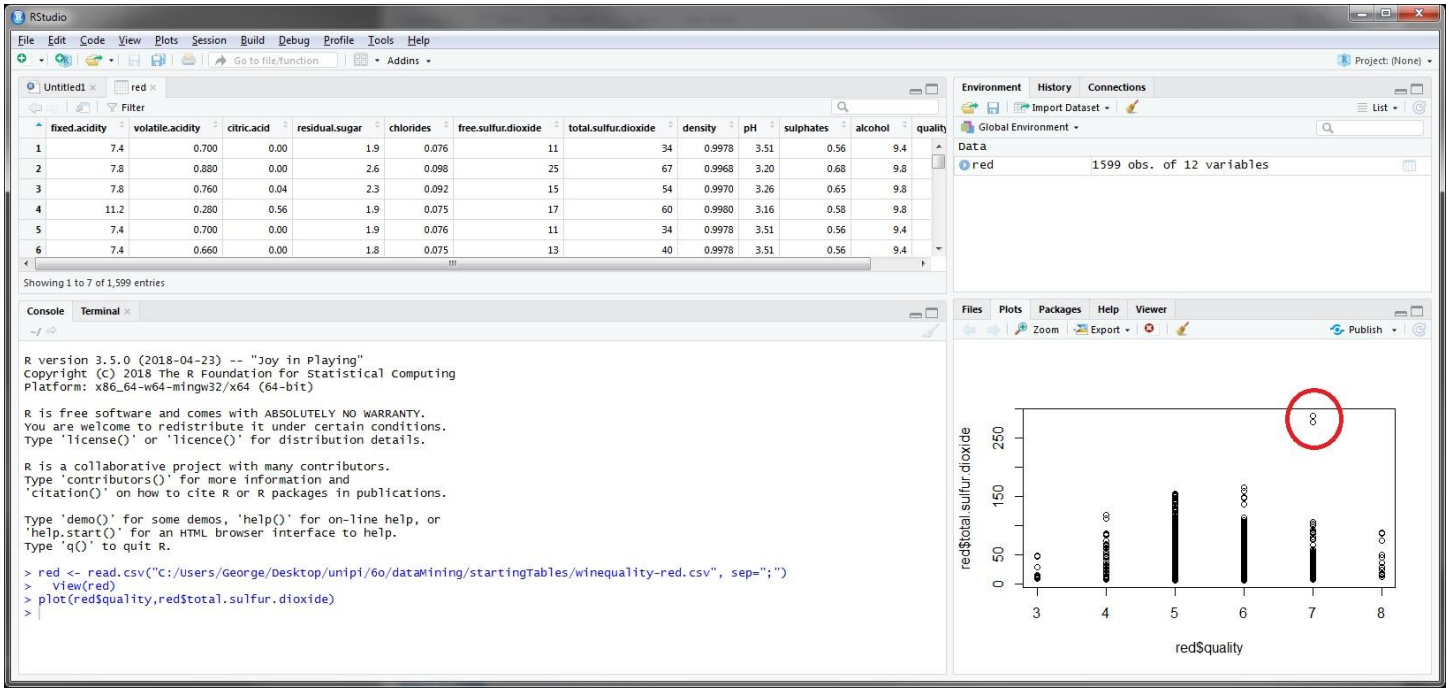


1ο βήμα

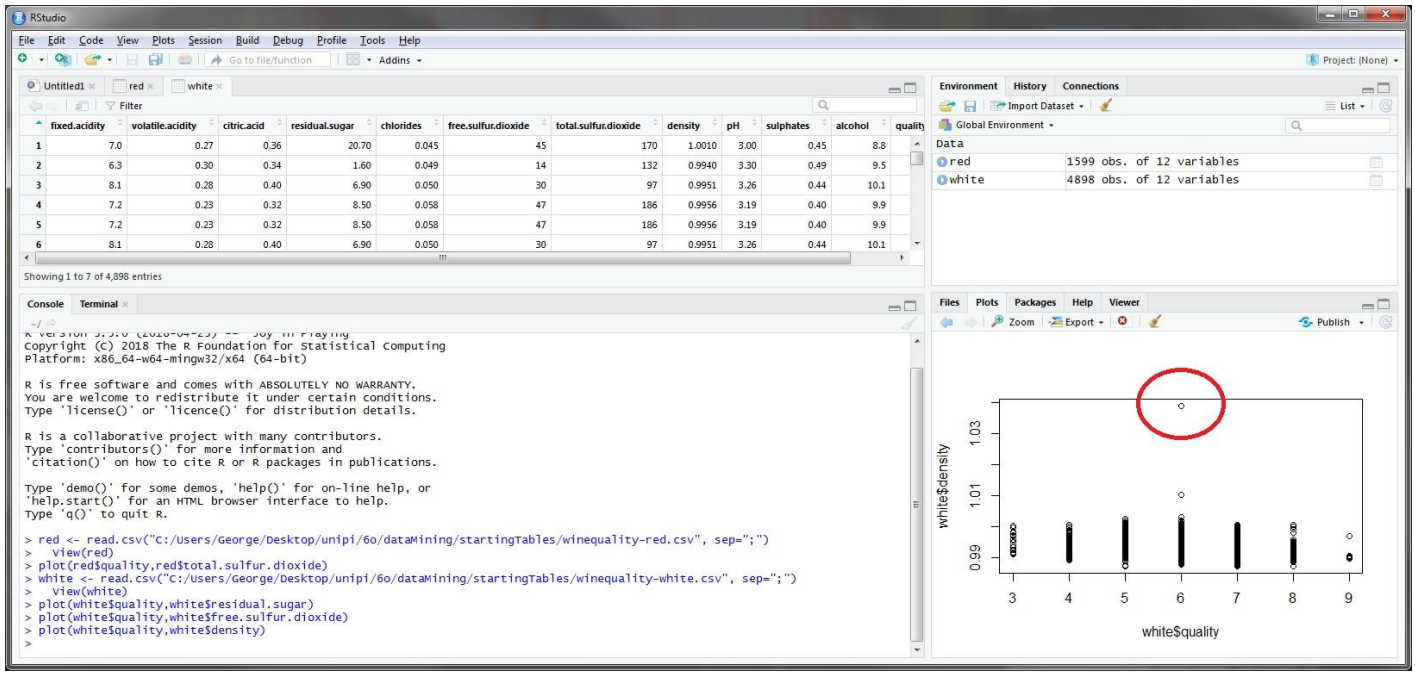
Δημιούργησα δύο πίνακες σε μια βάση δεδομένων υλοποιημένη σε PostgreSQL. Τα queries για τα create table υπάρχουν στον φάκελο createTables που θα βρείτε μέσα στο zip.

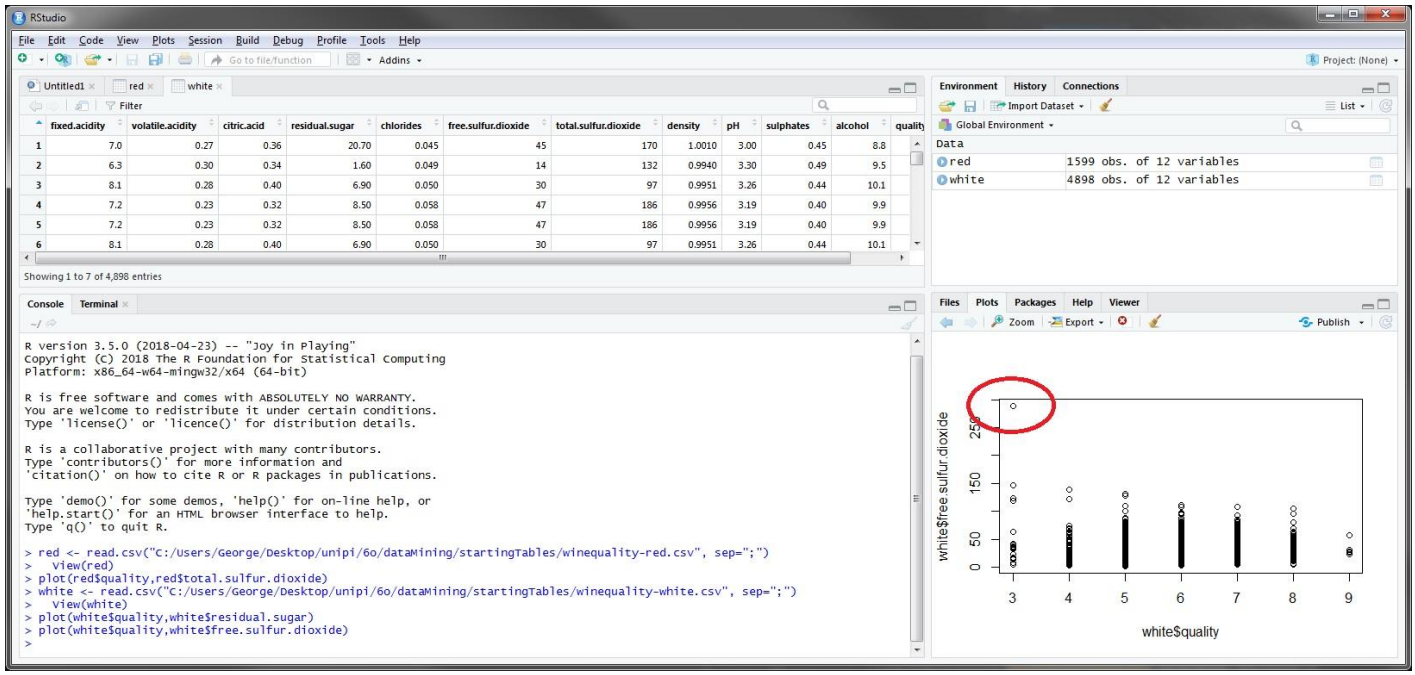
2ο βήμα

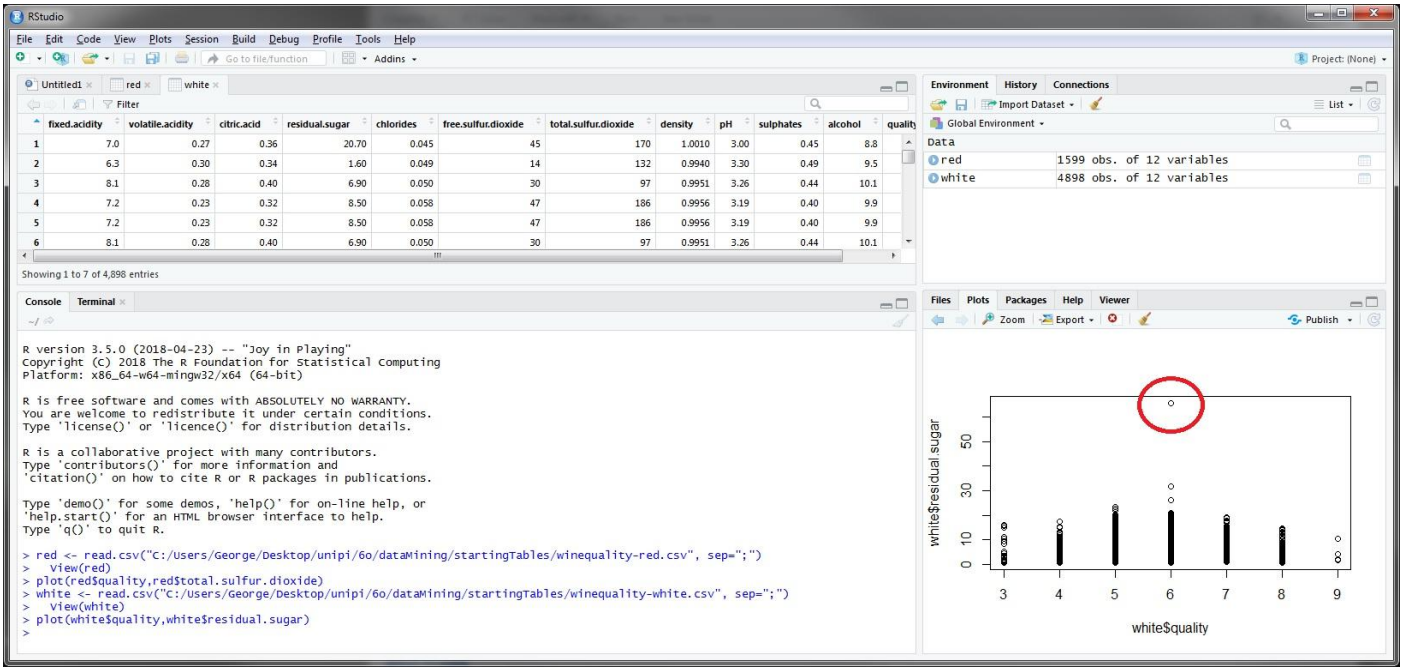
a) Αρχικά, φόρτωσα τα data στο R Studio. Έπειτα, με την εντολή plot βρήκα στον κάθε πίνακα κάποιες εντολές οι οποίες θεώρησα πως είχαν λανθασμένες τιμές κρίνοντας από το πόσο μεγάλες διαφορές είχαν από τις υπόλοιπες τιμές στις αντίστοιχες στήλες. Πιο συγκεκριμένα, στον πίνακα red :



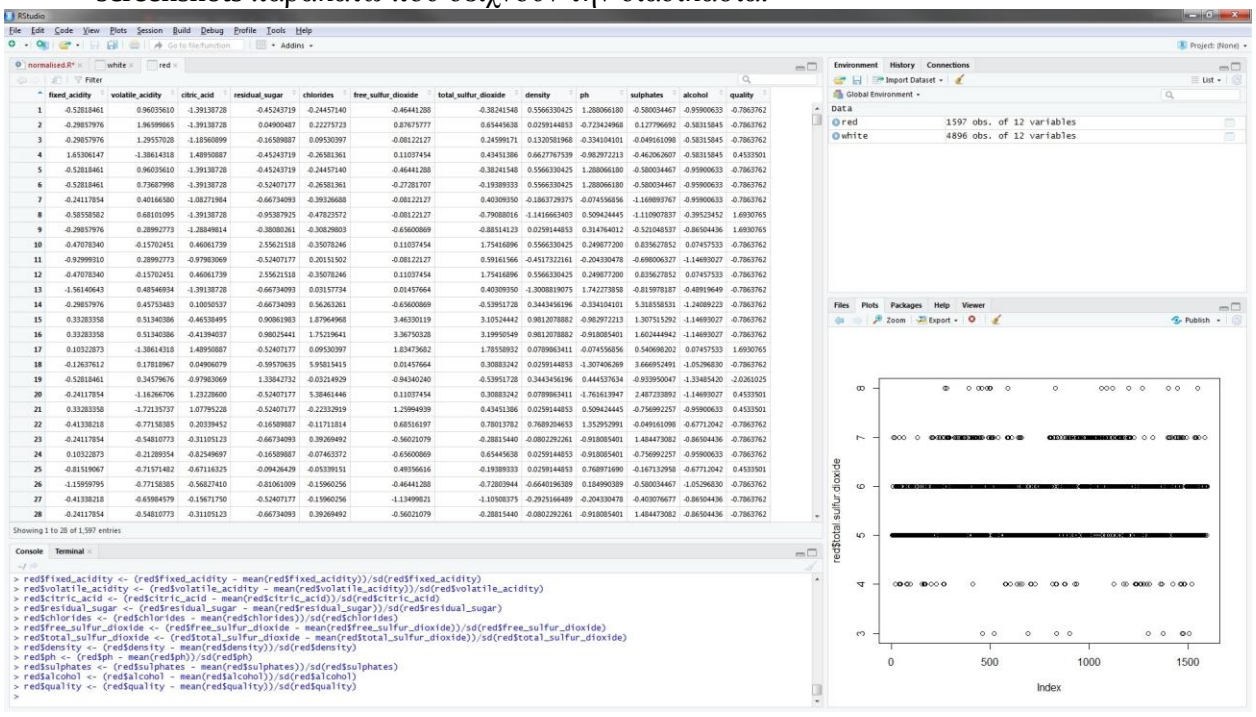
Και για τον white :

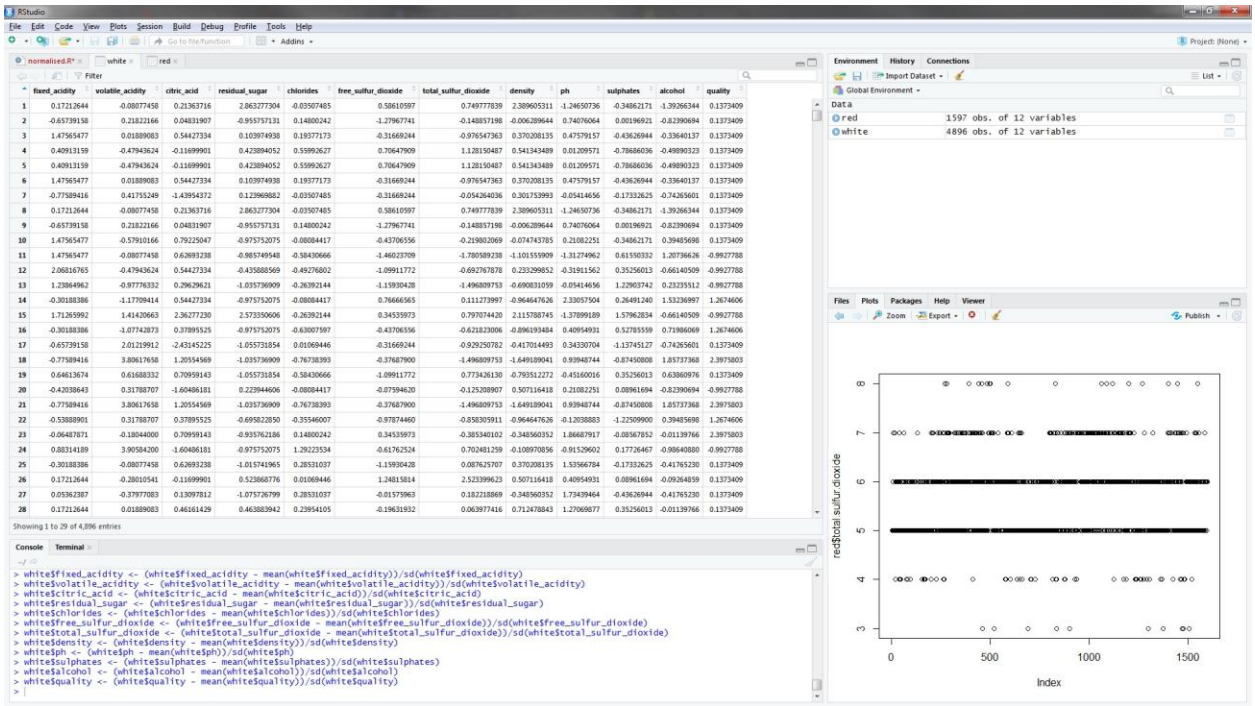






b) Στην συνέχεια, κανονικοποίησα τα δεδομένα. Αυτό το έκανα ουσιαστικά βρίσκοντας την μέση τιμή και την τυπική απόκλιση βάζοντας την μέση τιμή στο 0 και το έκανα με τον κώδικα που θα βρείτε στον φάκελο NormaliseFunc. Έχω επισυνάψει και μερικά screenshots παρακάτω που δείχνουν την διαδικασία:




RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
normalised.R* white red
Filter
fixed_acidity volatile_acidity citric_acid residual_sugar chlorides free_sulfur_dioxide total_sulfur_dioxide density ph sulphates alcohol quality
1 0.17212644 -0.08077458 0.21363716 2.863277304 -0.03507485 0.58610597 0.749777839 2.389605311 -1.24650736 -0.34862171 -1.39266344 0.1373409
2 -0.65739158 0.21822166 0.04831907 -0.955757131 0.14800242 -1.27967741 -0.148857198 -0.006289644 0.74076064 0.00196921 -0.82390694 0.1373409
3 1.47565477 0.01889083 0.54427334 0.103974938 0.19377173 -0.31669244 -0.976547363 0.370208135 0.47579157 -0.43626944 -0.33640137 0.1373409
4 0.40913159 -0.47943624 -0.11699901 0.423894052 0.55992627 0.70647909 1.128150487 0.541343489 0.01209571 -0.78686036 -0.49890323 0.1373409
5 0.40913159 -0.47943624 -0.11699901 0.423894052 0.55992627 0.70647909 1.128150487 0.541343489 0.01209571 -0.78686036 -0.49890323 0.1373409
6 1.47565477 0.01889083 0.54427334 0.103974938 0.19377173 -0.31669244 -0.976547363 0.370208135 0.47579157 -0.43626944 -0.33640137 0.1373409
7 -0.77589416 0.41755249 -1.43954372 0.123969882 -0.03507485 -0.31669244 -0.054264036 0.301753993 -0.05414656 -0.17332625 -0.74265601 0.1373409
8 0.17212644 -0.08077458 0.21363716 2.863277304 -0.03507485 0.58610597 0.749777839 2.389605311 -1.24650736 -0.34862171 -1.39266344 0.1373409
9 -0.65739158 0.21822166 0.04831907 -0.955757131 0.14800242 -1.27967741 -0.148857198 -0.006289644 0.74076064 0.00196921 -0.82390694 0.1373409
10 1.47565477 -0.57910166 0.79225047 -0.975752075 -0.08084417 -0.43706556 -0.219802069 -0.074743785 0.21082251 -0.34862171 0.39485698 0.1373409
11 1.47565477 -0.08077458 0.62693238 -0.985749548 -0.58430666 -1.46023709 -1.780589238 -1.101555909 -1.31274962 0.61550332 1.20736626 -0.9927788
12 2.06816765 -0.47943624 0.54427334 -0.435888569 -0.49276802 -1.09911772 -0.692767878 0.233299852 -0.31911562 0.35256013 -0.66140509 -0.9927788
13 1.23864962 -0.97776332 0.29629621 -1.035736909 -0.26392144 -1.15930428 -1.496809753 -0.690831059 -0.05414656 1.22903742 0.23235512 -0.9927788
14 -0.30188386 -1.17709414 0.54427334 -0.975752075 -0.08084417 0.76666565 0.111273997 -0.964647626 2.33057504 0.26491240 1.53236997 1.2674606
15 1.71265992 1.41420663 2.36277230 2.573350606 -0.26392144 0.34535973 0.797074420 2.115788745 -1.37899189 1.57962834 -0.66140509 -0.9927788
16 -0.30188386 -1.07742873 0.37895525 -0.975752075 -0.63007597 -0.43706556 -0.621823006 -0.896193484 0.40954931 0.52785559 0.71986069 1.2674606
17 -0.65739158 2.01219912 -2.43145225 -1.055731854 0.01069446 -0.31669244 -0.929250782 -0.417014493 0.34330704 -1.13745127 -0.74265601 0.1373409
18 -0.77589416 3.80617658 1.20554569 -1.035736909 -0.76738393 -0.37687900 -1.496809753 -1.649189041 0.93948744 -0.87450808 1.85737368 2.3975803
19 0.64613674 0.61688332 0.70959143 -1.055731854 -0.58430666 -1.09911772 0.773426130 -0.793512272 -0.45160016 0.35256013 0.63860976 0.1373409
20 -0.42038643 0.31788707 -1.60486181 0.223944606 -0.08084417 -0.07594620 -0.125208907 0.507116418 0.21082251 0.08961694 -0.82390694 -0.9927788
21 -0.77589416 3.80617658 1.20554569 -1.035736909 -0.76738393 -0.37687900 -1.496809753 -1.649189041 0.93948744 -0.87450808 1.85737368 2.3975803
22 -0.53888901 0.31788707 0.37895525 -0.695822850 -0.35546007 -0.97874460 -0.858305911 -0.964647626 -0.12038883 -1.22509900 0.39485698 1.2674606
23 -0.06487871 -0.18044000 0.70959143 -0.935762186 0.14800242 0.34535973 -0.385340102 -0.348560352 1.86687917 -0.08567852 -0.01139766 2.3975803
24 0.88314189 3.90584200 -1.60486181 -0.975752075 1.29223534 -0.61762524 0.702481259 -0.108970856 -0.91529602 0.17726467 -0.98640880 -0.9927788
25 -0.30188386 -0.08077458 0.62693238 -1.015741965 0.28531037 -1.15930428 0.087625707 0.370208135 1.53566784 -0.17332625 -0.41765230 0.1373409
26 0.17212644 -0.28010541 -0.11699901 0.523868776 0.01069446 1.24815814 2.523399623 0.507116418 0.40954931 0.08961694 -0.09264859 0.1373409
27 0.05362387 -0.37977083 0.13097812 -1.075726799 0.28531037 -0.01575963 0.182218869 -0.348560352 1.73439464 -0.43626944 -0.41765230 0.1373409
28 0.17212644 0.01889083 0.46161429 0.463883942 0.23954105 -0.19631932 0.063977416 0.712478843 1.27069877 0.35256013 -0.01139766 0.1373409
Showing 1 to 29 of 4,896 entries
Console Terminal
> white$fixed_acidity <- (white$fixed_acidity - mean(white$fixed_acidity))/sd(white$fixed_acidity)
> white$volatile_acidity <- (white$volatile_acidity - mean(white$volatile_acidity))/sd(white$volatile_acidity)
> white$citric_acid <- (white$citric_acid - mean(white$citric_acid))/sd(white$citric_acid)
> white$residual_sugar <- (white$residual_sugar - mean(white$residual_sugar))/sd(white$residual_sugar)
> white$chlorides <- (white$chlorides - mean(white$chlorides))/sd(white$chlorides)
> white$free_sulfur_dioxide <- (white$free_sulfur_dioxide - mean(white$free_sulfur_dioxide))/sd(white$free_sulfur_dioxide)
> white$total_sulfur_dioxide <- (white$total_sulfur_dioxide - mean(white$total_sulfur_dioxide))/sd(white$total_sulfur_dioxide)
> white$density <- (white$density - mean(white$density))/sd(white$density)
> white$ph <- (white$ph - mean(white$ph))/sd(white$ph)
> white$sulphates <- (white$sulphates - mean(white$sulphates))/sd(white$sulphates)
> white$alcohol <- (white$alcohol - mean(white$alcohol))/sd(white$alcohol)
> white$quality <- (white$quality - mean(white$quality))/sd(white$quality)
Environment History Connections
Import Dataset List
Global Environment
Data
red 1597 obs. of 12 variables
white 4896 obs. of 12 variables
Files Plots Packages Help Viewer
Zoom Export Publish
red$total.sulfur.dioxide
Index
0 500 1000 1500
3 4 5 6 7 8

c) Σε αυτό το βήμα χρησιμοποίησα την εντολή cor ανάμεσα στην στήλη quality, που είναι άλλωστε και η στήλη στόχος, και σε κάθε άλλη στήλη στον πίνακα και πήρα τα παρακάτω αποτελέσματα. Γνωρίζοντας ότι όσο πιο κοντά είναι το correlation στο 1 ή στο -1 τόσο πιο γραμμικώς εξαρτημένα είναι τα δύο μεγέθη. Αν είναι θετικό σημαίνει ότι είναι ανάλογα και αν είναι αρνητικό αντιστρόφως ανάλογα. Συνεπώς μπορούμε να ποιες στήλες είναι πιο κοντά στο 0 δεν συνεισφέρουν πολύ για την πρόβλεψη του quality οπότε θα μπορούσα να τα παραλείψω σε επόμενα βήματα.