VytKra Mathematical Methods for Artificial Intelligence Lab 1

# Reading Data

data <- read.csv("superconductor\_dataset.csv")

# EDA

library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(SmartEDA)

## Warning: package 'SmartEDA' was built under R version 4.0.5

## Registered S3 method overwritten by 'GGally':  
## method from   
## +.gg ggplot2

# Overview of the data - Type = 1  
ExpData(data,type=1)

## Descriptions Value  
## 1 Sample size (nrow) 21263  
## 2 No. of variables (ncol) 169  
## 3 No. of numeric/interger variables 168  
## 4 No. of factor variables 0  
## 5 No. of text variables 1  
## 6 No. of logical variables 0  
## 7 No. of identifier variables 0  
## 8 No. of date variables 0  
## 9 No. of zero variance variables (uniform) 9  
## 10 %. of variables having complete cases 100% (169)  
## 11 %. of variables having >0% and <50% missing cases 0% (0)  
## 12 %. of variables having >=50% and <90% missing cases 0% (0)  
## 13 %. of variables having >=90% missing cases 0% (0)

# we have only 1 text variable, no missing values  
  
n\_distinct(data %>% select(material))

## [1] 15542

nrow(data)

## [1] 21263

# 15542 unique values out of 21263, we will not use this column  
  
data <- data %>% select(-material)  
  
  
library(purrr)  
  
# Return a character vector of variable names which have 0 variance  
variables\_with\_zero\_var <- names(data)[which(map\_dbl(data, var) == 0)]  
  
summary(data[,variables\_with\_zero\_var])

## He Ne Ar Kr Xe Pm   
## Min. :0 Min. :0 Min. :0 Min. :0 Min. :0 Min. :0   
## 1st Qu.:0 1st Qu.:0 1st Qu.:0 1st Qu.:0 1st Qu.:0 1st Qu.:0   
## Median :0 Median :0 Median :0 Median :0 Median :0 Median :0   
## Mean :0 Mean :0 Mean :0 Mean :0 Mean :0 Mean :0   
## 3rd Qu.:0 3rd Qu.:0 3rd Qu.:0 3rd Qu.:0 3rd Qu.:0 3rd Qu.:0   
## Max. :0 Max. :0 Max. :0 Max. :0 Max. :0 Max. :0   
## Po At Rn   
## Min. :0 Min. :0 Min. :0   
## 1st Qu.:0 1st Qu.:0 1st Qu.:0   
## Median :0 Median :0 Median :0   
## Mean :0 Mean :0 Mean :0   
## 3rd Qu.:0 3rd Qu.:0 3rd Qu.:0   
## Max. :0 Max. :0 Max. :0

# Those columns dont hold any info, we'll exclude them  
  
data <- data %>% select(-all\_of(variables\_with\_zero\_var))  
  
  
# Target variable  
summary( data %>% select(critical\_temp) )

## critical\_temp   
## Min. : 0.00021   
## 1st Qu.: 5.36500   
## Median : 20.00000   
## Mean : 34.42122   
## 3rd Qu.: 63.00000   
## Max. :185.00000

# Correlation  
library(corrplot)

## Warning: package 'corrplot' was built under R version 4.0.5

## corrplot 0.84 loaded

corr\_simple <- function(df,sig=0.5){  
 corr <- cor(df)  
 #prepare to drop duplicates and correlations of 1   
 corr[lower.tri(corr,diag=TRUE)] <- NA   
 #drop perfect correlations  
 corr[corr == 1] <- NA   
 #turn into a 3-column table  
 corr <- as.data.frame(as.table(corr))  
 #remove the NA values from above   
 corr <- na.omit(corr)   
 #select significant values   
 corr <- subset(corr, abs(Freq) > sig)   
 #sort by highest correlation  
 corr <- corr[order(-abs(corr$Freq)),]   
 return(corr)  
}  
  
  
ExpNumStat(data,by ="A",Outlier=TRUE,round= 2, gp = "critical\_temp") %>%   
 select(Vname, nNeg, nPos, nZero, min, max, mean, median, SD, Skewness, nOutliers, cor)

## Note: Target variable is continuous  
## Summary statistics excluded group by statement  
## Results generated with correlation value against target variable

## Vname nNeg nPos nZero min max mean  
## 123 Ag 0 156 21107 0.00 7.00 0.01  
## 91 Al 0 731 20532 0.00 99.92 0.06  
## 110 As 0 1502 19761 0.00 18.00 0.16  
## 153 Au 0 242 21021 0.00 64.00 0.02  
## 84 B 0 1205 20058 0.00 105.00 0.14  
## 131 Ba 0 6751 14512 0.00 24.00 0.57  
## 83 Be 0 96 21167 0.00 40.00 0.03  
## 157 Bi 0 2389 18874 0.00 14.00 0.20  
## 112 Br 0 66 21197 0.00 5.00 0.00  
## 85 C 0 1274 19989 0.00 120.00 0.38  
## 97 Ca 0 4112 17151 0.00 24.00 0.26  
## 124 Cd 0 178 21085 0.00 100.00 0.01  
## 133 Ce 0 1162 20101 0.00 5.00 0.03  
## 95 Cl 0 146 21117 0.00 3.00 0.01  
## 104 Co 0 1035 20228 0.00 35.38 0.04  
## 101 Cr 0 195 21068 0.00 34.90 0.01  
## 80 critical\_temp 0 21263 0 0.00 185.00 34.42  
## 130 Cs 0 98 21165 0.00 3.00 0.00  
## 106 Cu 0 10838 10425 0.00 98.00 1.28  
## 140 Dy 0 239 21024 0.00 5.00 0.01  
## 5 entropy\_atomic\_mass 0 20978 285 0.00 1.98 1.17  
## 25 entropy\_atomic\_radius 0 20978 285 0.00 2.14 1.27  
## 35 entropy\_Density 0 20978 285 0.00 1.95 1.07  
## 45 entropy\_ElectronAffinity 0 20978 285 0.00 1.77 1.07  
## 15 entropy\_fie 0 20978 285 0.00 2.16 1.30  
## 55 entropy\_FusionHeat 0 20978 285 0.00 2.03 1.09  
## 65 entropy\_ThermalConductivity 0 20978 285 0.00 1.63 0.73  
## 75 entropy\_Valence 0 20978 285 0.00 2.14 1.30  
## 142 Er 0 356 20907 0.00 5.00 0.01  
## 137 Eu 0 380 20883 0.00 6.00 0.02  
## 88 F 0 669 20594 0.00 4.00 0.01  
## 103 Fe 0 2339 18924 0.00 30.00 0.15  
## 108 Ga 0 605 20658 0.00 41.00 0.07  
## 138 Gd 0 663 20600 0.00 4.00 0.02  
## 109 Ge 0 520 20743 0.00 46.00 0.08  
## 3 gmean\_atomic\_mass 0 21263 0 5.32 208.98 71.29  
## 23 gmean\_atomic\_radius 0 21263 0 48.00 298.00 144.45  
## 33 gmean\_Density 0 21263 0 1.43 22590.00 3460.69  
## 43 gmean\_ElectronAffinity 0 21263 0 1.50 326.10 54.36  
## 13 gmean\_fie 0 21263 0 375.50 1313.10 737.47  
## 53 gmean\_FusionHeat 0 21263 0 0.22 105.00 10.14  
## 63 gmean\_ThermalConductivity 0 21263 0 0.03 317.88 29.84  
## 73 gmean\_Valence 0 21263 0 1.00 7.00 3.06  
## 81 H 0 299 20964 0.00 14.00 0.02  
## 146 Hf 0 222 21041 0.00 25.00 0.01  
## 154 Hg 0 845 20418 0.00 8.00 0.04  
## 141 Ho 0 227 21036 0.00 5.00 0.01  
## 129 I 0 83 21180 0.00 4.00 0.00  
## 115 Y 0 4075 17188 0.00 9.00 0.18  
## 144 Yb 0 291 20972 0.00 16.00 0.01  
## 125 In 0 544 20719 0.00 31.50 0.05  
## 151 Ir 0 572 20691 0.00 45.00 0.06  
## 96 K 0 530 20733 0.00 3.30 0.02  
## 132 La 0 3463 17800 0.00 98.00 0.26  
## 82 Li 0 311 20952 0.00 3.00 0.01  
## 145 Lu 0 399 20864 0.00 7.00 0.03  
## 1 mean\_atomic\_mass 0 21263 0 6.94 208.98 87.56  
## 21 mean\_atomic\_radius 0 21263 0 48.00 298.00 157.98  
## 31 mean\_Density 0 21263 0 1.43 22590.00 6111.47  
## 41 mean\_ElectronAffinity 0 21263 0 1.50 326.10 76.88  
## 11 mean\_fie 0 21263 0 375.50 1313.10 769.61  
## 51 mean\_FusionHeat 0 21263 0 0.22 105.00 14.30  
## 61 mean\_ThermalConductivity 0 21263 0 0.03 332.50 89.71  
## 71 mean\_Valence 0 21263 0 1.00 7.00 3.20  
## 90 Mg 0 522 20741 0.00 12.00 0.03  
## 102 Mn 0 171 21092 0.00 14.00 0.00  
## 118 Mo 0 888 20375 0.00 99.99 0.15  
## 86 N 0 306 20957 0.00 12.80 0.01  
## 89 Na 0 322 20941 0.00 4.00 0.01  
## 117 Nb 0 1436 19827 0.00 99.98 0.44  
## 135 Nd 0 946 20317 0.00 6.00 0.04  
## 105 Ni 0 1149 20114 0.00 45.00 0.09  
## 87 O 0 11964 9299 0.00 66.00 3.01  
## 150 Os 0 255 21008 0.00 10.00 0.02  
## 93 P 0 355 20908 0.00 20.00 0.03  
## 156 Pb 0 1255 20008 0.00 19.00 0.04  
## 122 Pd 0 487 20776 0.00 51.00 0.09  
## 134 Pr 0 1276 19987 0.00 185.00 0.04  
## 152 Pt 0 419 20844 0.00 5.80 0.03  
## 7 range\_atomic\_mass 0 20978 285 0.00 207.97 115.60  
## 27 range\_atomic\_radius 0 20958 305 0.00 256.00 139.33  
## 37 range\_Density 0 20964 299 0.00 22588.57 8665.44  
## 47 range\_ElectronAffinity 0 20953 310 0.00 349.00 120.73  
## 17 range\_fie 0 20974 289 0.00 1304.50 572.22  
## 57 range\_FusionHeat 0 20978 285 0.00 104.78 21.14  
## 67 range\_ThermalConductivity 0 20962 301 0.00 429.97 250.89  
## 113 Rb 0 158 21105 0.00 4.00 0.01  
## 149 Re 0 360 20903 0.00 97.24 0.04  
## 121 Rh 0 643 20620 0.00 45.00 0.07  
## 120 Ru 0 735 20528 0.00 64.00 0.06  
## 94 S 0 694 20569 0.00 15.00 0.11  
## 127 Sb 0 343 20920 0.00 83.50 0.10  
## 98 Sc 0 149 21114 0.00 5.00 0.01  
## 111 Se 0 685 20578 0.00 19.00 0.08  
## 92 Si 0 725 20538 0.00 100.00 0.19  
## 136 Sm 0 478 20785 0.00 12.00 0.02  
## 126 Sn 0 840 20423 0.00 99.20 0.12  
## 114 Sr 0 4852 16411 0.00 16.70 0.33  
## 9 std\_atomic\_mass 0 20978 285 0.00 101.02 44.39  
## 29 std\_atomic\_radius 0 20958 305 0.00 115.50 51.60  
## 39 std\_Density 0 20964 299 0.00 10724.37 3416.91  
## 49 std\_ElectronAffinity 0 20953 310 0.00 162.90 48.91  
## 19 std\_fie 0 20974 289 0.00 499.67 215.63  
## 59 std\_FusionHeat 0 20978 285 0.00 51.63 8.32  
## 69 std\_ThermalConductivity 0 20962 301 0.00 214.99 98.94  
## 78 std\_Valence 0 19865 1398 0.00 3.00 0.84  
## 147 Ta 0 396 20867 0.00 55.00 0.04  
## 139 Tb 0 104 21159 0.00 5.00 0.00  
## 119 Tc 0 50 21213 0.00 6.00 0.00  
## 128 Te 0 527 20736 0.00 66.70 0.04  
## 99 Ti 0 589 20674 0.00 75.00 0.16  
## 155 Tl 0 908 20355 0.00 7.00 0.05  
## 143 Tm 0 181 21082 0.00 5.00 0.01  
## 100 V 0 798 20465 0.00 79.50 0.22  
## 148 W 0 265 20998 0.00 14.00 0.01  
## 6 wtd\_entropy\_atomic\_mass 0 20978 285 0.00 1.96 1.06  
## 26 wtd\_entropy\_atomic\_radius 0 20978 285 0.00 1.90 1.13  
## 36 wtd\_entropy\_Density 0 20978 285 0.00 1.70 0.86  
## 46 wtd\_entropy\_ElectronAffinity 0 20978 285 0.00 1.68 0.77  
## 16 wtd\_entropy\_fie 0 20978 285 0.00 2.04 0.93  
## 56 wtd\_entropy\_FusionHeat 0 20978 285 0.00 1.75 0.91  
## 66 wtd\_entropy\_ThermalConductivity 0 20978 285 0.00 1.61 0.54  
## 76 wtd\_entropy\_Valence 0 20978 285 0.00 1.95 1.05  
## 4 wtd\_gmean\_atomic\_mass 0 21263 0 1.96 208.98 58.54  
## 24 wtd\_gmean\_atomic\_radius 0 21263 0 48.00 298.00 120.99  
## 34 wtd\_gmean\_Density 0 21263 0 0.69 22590.00 3117.24  
## 44 wtd\_gmean\_ElectronAffinity 0 21263 0 1.50 326.10 72.42  
## 14 wtd\_gmean\_fie 0 21263 0 375.50 1327.59 832.77  
## 54 wtd\_gmean\_FusionHeat 0 21263 0 0.22 105.00 10.14  
## 64 wtd\_gmean\_ThermalConductivity 0 21263 0 0.02 376.03 27.31  
## 74 wtd\_gmean\_Valence 0 21263 0 1.00 7.00 3.06  
## 2 wtd\_mean\_atomic\_mass 0 21263 0 6.42 208.98 72.99  
## 22 wtd\_mean\_atomic\_radius 0 21263 0 48.00 298.00 134.72  
## 32 wtd\_mean\_Density 0 21263 0 1.43 22590.00 5267.19  
## 42 wtd\_mean\_ElectronAffinity 0 21263 0 1.50 326.10 92.72  
## 12 wtd\_mean\_fie 0 21263 0 375.50 1348.03 870.44  
## 52 wtd\_mean\_FusionHeat 0 21263 0 0.22 105.00 13.85  
## 62 wtd\_mean\_ThermalConductivity 0 21263 0 0.03 406.96 81.55  
## 72 wtd\_mean\_Valence 0 21263 0 1.00 7.00 3.15  
## 8 wtd\_range\_atomic\_mass 0 20978 285 0.00 205.59 33.23  
## 28 wtd\_range\_atomic\_radius 0 20977 286 0.00 240.16 51.37  
## 38 wtd\_range\_Density 0 20976 287 0.00 22434.16 2902.74  
## 48 wtd\_range\_ElectronAffinity 0 20972 291 0.00 218.70 59.33  
## 18 wtd\_range\_fie 0 20978 285 0.00 1251.86 483.52  
## 58 wtd\_range\_FusionHeat 0 20978 285 0.00 102.67 8.22  
## 68 wtd\_range\_ThermalConductivity 0 20971 292 0.00 401.44 62.03  
## 77 wtd\_range\_Valence 0 20886 377 0.00 6.99 1.48  
## 10 wtd\_std\_atomic\_mass 0 20978 285 0.00 101.02 41.45  
## 30 wtd\_std\_atomic\_radius 0 20958 305 0.00 97.14 52.34  
## 40 wtd\_std\_Density 0 20965 298 0.00 10410.93 3319.17  
## 50 wtd\_std\_ElectronAffinity 0 20959 304 0.00 169.08 44.41  
## 20 wtd\_std\_fie 0 20975 288 0.00 479.16 224.05  
## 60 wtd\_std\_FusionHeat 0 20978 285 0.00 51.68 7.72  
## 70 wtd\_std\_ThermalConductivity 0 20963 300 0.00 213.30 96.23  
## 79 wtd\_std\_Valence 0 19996 1267 0.00 3.00 0.67  
## 107 Zn 0 832 20431 0.00 20.00 0.01  
## 116 Zr 0 631 20632 0.00 96.71 0.37  
## median SD Skewness nOutliers cor  
## 123 0.00 0.17 35.04 156 -0.03  
## 91 0.00 1.13 47.38 731 -0.04  
## 110 0.00 1.08 14.30 1502 -0.06  
## 153 0.00 0.72 73.67 242 -0.03  
## 84 0.00 1.04 51.63 1205 -0.09  
## 131 0.00 0.98 3.88 167 0.56  
## 83 0.00 0.85 29.38 96 -0.03  
## 157 0.00 0.66 4.91 2389 0.16  
## 112 0.00 0.08 28.36 66 -0.03  
## 85 0.00 4.41 14.00 1274 -0.04  
## 97 0.00 0.90 7.74 4112 0.30  
## 124 0.00 0.69 143.92 178 -0.01  
## 133 0.00 0.17 8.18 1162 -0.10  
## 95 0.00 0.12 14.98 146 -0.03  
## 104 0.00 0.58 44.07 1035 -0.04  
## 101 0.00 0.25 122.79 195 -0.01  
## 80 20.00 34.25 0.86 1 1.00  
## 130 0.00 0.08 26.30 98 -0.02  
## 106 0.90 2.08 12.97 72 0.52  
## 140 0.00 0.10 15.82 239 0.02  
## 5 1.20 0.36 -0.79 387 0.54  
## 25 1.33 0.38 -0.75 285 0.56  
## 35 1.09 0.34 -0.81 602 0.46  
## 45 1.14 0.34 -0.91 567 0.44  
## 15 1.36 0.38 -0.79 285 0.57  
## 55 1.11 0.38 -0.58 285 0.55  
## 65 0.74 0.33 -0.12 0 0.09  
## 75 1.37 0.39 -0.77 285 0.60  
## 142 0.00 0.13 15.44 356 0.04  
## 137 0.00 0.15 11.48 380 0.02  
## 88 0.00 0.13 15.02 669 0.03  
## 103 0.00 0.71 11.72 2339 -0.08  
## 108 0.00 1.12 20.92 605 -0.04  
## 138 0.00 0.16 7.83 663 0.08  
## 109 0.00 1.02 20.15 520 -0.07  
## 3 66.36 31.03 1.47 3314 -0.23  
## 23 142.81 22.09 0.09 1503 -0.14  
## 33 1339.97 3703.26 1.66 494 -0.54  
## 43 51.47 29.01 1.35 822 -0.38  
## 13 727.96 78.33 1.19 1480 -0.03  
## 53 5.25 10.07 2.67 1485 -0.43  
## 63 14.29 34.06 2.34 1245 -0.39  
## 73 2.62 1.05 1.17 494 -0.57  
## 81 0.00 0.27 33.00 299 -0.03  
## 146 0.00 0.21 87.98 222 -0.03  
## 154 0.00 0.21 9.45 845 0.22  
## 141 0.00 0.10 17.28 227 0.01  
## 129 0.00 0.09 24.74 83 -0.01  
## 115 0.00 0.43 4.35 4075 0.25  
## 144 0.00 0.21 54.54 291 0.03  
## 125 0.00 0.52 22.53 544 -0.09  
## 151 0.00 0.86 38.09 572 -0.06  
## 96 0.00 0.14 14.20 530 -0.05  
## 132 0.00 2.32 38.52 3463 -0.05  
## 82 0.00 0.13 12.97 311 -0.05  
## 145 0.00 0.28 14.72 399 -0.06  
## 1 84.92 29.68 0.77 1606 -0.11  
## 21 160.25 20.15 -0.57 1203 0.11  
## 31 5329.09 2846.79 2.10 1828 -0.37  
## 41 73.10 27.70 1.10 2225 -0.19  
## 11 764.90 87.49 1.01 1913 0.10  
## 51 9.30 11.30 2.52 1562 -0.39  
## 61 96.50 38.52 0.22 300 0.38  
## 71 2.83 1.04 1.00 7 -0.60  
## 90 0.00 0.27 28.26 522 -0.03  
## 102 0.00 0.13 74.33 171 -0.01  
## 118 0.00 2.08 37.96 888 -0.06  
## 86 0.00 0.15 36.57 306 -0.05  
## 89 0.00 0.10 17.69 322 -0.05  
## 117 0.00 4.85 15.30 1436 -0.06  
## 135 0.00 0.22 6.98 946 0.00  
## 105 0.00 0.98 26.13 1149 -0.07  
## 87 1.00 3.81 2.26 50 0.57  
## 150 0.00 0.28 16.23 255 -0.07  
## 93 0.00 0.47 29.51 355 -0.05  
## 156 0.00 0.27 25.83 1255 0.02  
## 122 0.00 1.55 28.80 487 -0.05  
## 134 0.00 1.28 141.25 1276 -0.01  
## 152 0.00 0.31 10.44 419 -0.08  
## 7 122.91 54.63 -0.38 0 0.49  
## 27 171.00 67.27 -0.69 0 0.65  
## 37 8958.57 4097.13 0.38 2908 0.26  
## 47 127.05 58.70 0.94 1524 0.28  
## 17 764.10 309.61 -0.46 0 0.60  
## 57 12.88 20.37 2.75 2112 -0.14  
## 67 399.80 158.70 -0.23 0 0.69  
## 113 0.00 0.12 20.43 158 -0.03  
## 149 0.00 1.18 62.61 360 -0.03  
## 121 0.00 1.01 34.14 643 -0.06  
## 120 0.00 0.77 49.71 735 -0.05  
## 94 0.00 0.76 9.36 694 -0.11  
## 127 0.00 1.84 28.44 343 -0.05  
## 98 0.00 0.19 21.65 149 -0.04  
## 111 0.00 0.68 15.17 685 -0.09  
## 92 0.00 2.22 21.68 725 -0.07  
## 136 0.00 0.18 20.62 478 0.01  
## 126 0.00 1.89 35.58 840 -0.06  
## 114 0.00 0.76 3.41 4852 0.27  
## 9 45.12 20.04 -0.35 4 0.38  
## 29 58.66 22.90 -0.75 0 0.56  
## 39 3301.89 1673.62 0.86 3497 0.12  
## 49 51.13 21.74 0.58 2519 0.26  
## 19 266.37 109.97 -0.45 0 0.54  
## 59 4.95 8.67 2.82 2090 -0.20  
## 69 135.76 60.14 -0.23 0 0.65  
## 78 0.80 0.48 0.44 44 -0.21  
## 147 0.00 0.85 47.42 396 -0.04  
## 139 0.00 0.06 43.10 104 -0.02  
## 119 0.00 0.06 52.60 50 -0.03  
## 128 0.00 0.72 76.87 527 -0.05  
## 99 0.00 2.73 22.01 589 -0.05  
## 155 0.00 0.27 7.40 908 0.17  
## 143 0.00 0.13 24.86 181 0.00  
## 100 0.00 3.41 19.30 798 -0.05  
## 148 0.00 0.16 43.56 265 -0.05  
## 6 1.15 0.40 -0.63 0 0.63  
## 26 1.24 0.41 -0.79 0 0.60  
## 36 0.88 0.32 -0.60 738 0.40  
## 46 0.78 0.29 -0.23 3428 0.24  
## 16 0.92 0.33 -0.03 2176 0.39  
## 56 0.99 0.37 -0.57 0 0.56  
## 66 0.55 0.32 0.31 3 -0.12  
## 76 1.17 0.38 -0.73 0 0.59  
## 4 39.92 36.65 1.70 1247 -0.37  
## 24 113.18 35.84 0.55 8 -0.41  
## 34 1515.36 3975.12 1.65 439 -0.54  
## 44 73.17 31.65 0.41 348 -0.11  
## 14 856.20 119.77 -0.31 4 0.34  
## 54 4.93 13.13 2.49 577 -0.43  
## 64 6.10 40.19 2.59 863 -0.37  
## 74 2.43 1.17 1.02 52 -0.62  
## 2 60.70 33.49 1.50 1239 -0.31  
## 22 125.97 28.80 0.57 11 -0.30  
## 32 4303.42 3221.31 2.14 926 -0.43  
## 42 102.86 32.28 -0.22 528 0.11  
## 12 889.97 143.28 -0.29 0 0.40  
## 52 8.33 14.28 2.76 1173 -0.39  
## 62 73.33 45.52 1.38 1420 0.38  
## 72 2.62 1.19 0.89 30 -0.63  
## 8 26.64 26.97 2.46 1629 -0.34  
## 28 43.00 35.02 1.64 1879 -0.34  
## 38 2082.96 2398.47 3.32 1657 -0.28  
## 48 71.16 28.62 0.11 170 0.19  
## 18 510.44 224.04 -0.28 0 0.30  
## 58 3.44 11.41 4.23 1374 -0.31  
## 68 56.56 43.12 1.42 196 0.47  
## 77 1.06 0.98 1.55 1394 -0.44  
## 10 44.29 19.98 -0.31 15 0.36  
## 30 59.93 25.29 -0.65 0 0.60  
## 40 3625.63 1611.80 0.45 2261 0.21  
## 50 48.03 20.43 0.52 1460 0.32  
## 20 258.45 127.93 -0.41 0 0.58  
## 60 5.50 7.29 2.85 2641 -0.20  
## 70 113.56 63.71 -0.09 0 0.72  
## 79 0.50 0.46 0.66 25 -0.30  
## 107 0.00 0.40 46.96 832 -0.02  
## 116 0.00 4.85 14.87 631 -0.07

# Data split

set.seed(123)  
library(caret)

## Warning: package 'caret' was built under R version 4.0.5

## Loading required package: lattice

## Loading required package: ggplot2

## Warning: package 'ggplot2' was built under R version 4.0.5

##   
## Attaching package: 'caret'

## The following object is masked from 'package:purrr':  
##   
## lift

indices <- createDataPartition(data$critical\_temp, p = 0.8, list = FALSE)  
train <- data[indices,]  
test <- data[-indices,]  
  
preProcValues <- preProcess(train, method = c("center", "scale"))  
train\_transformed <- predict(preProcValues, train)  
test\_transformed <- predict(preProcValues, test)  
  
label\_index <- which(colnames(train\_transformed) == "critical\_temp")

# Linear Regression

# linear regression  
train\_control\_cv <- trainControl(method = "cv", number = 10)  
fit\_lm <- train(critical\_temp ~ ., data = train\_transformed, method = "lm", trControl = train\_control\_cv)  
print(fit\_lm)

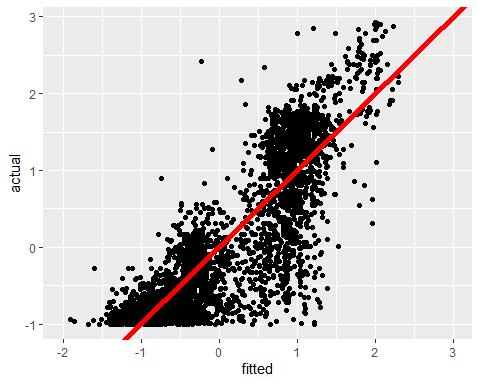
## Linear Regression   
##   
## 17011 samples  
## 158 predictor  
##   
## No pre-processing  
## Resampling: Cross-Validated (10 fold)   
## Summary of sample sizes: 15310, 15309, 15309, 15310, 15310, 15311, ...   
## Resampling results:  
##   
## RMSE Rsquared MAE   
## 0.5462596 0.7237787 0.3622424  
##   
## Tuning parameter 'intercept' was held constant at a value of TRUE

test\_lm <- predict(fit\_lm, newdata = test\_transformed)  
round(postResample(pred = test\_lm, obs = test\_transformed$critical\_temp), 3)

## RMSE Rsquared MAE   
## 1.209 0.288 0.382

fitted\_actual\_lm <- data.frame(fitted = test\_lm, actual = test\_transformed$critical\_temp)  
  
ggplot(fitted\_actual\_lm,   
 aes(x = fitted,  
 y = actual)) +  
 geom\_point() +  
 xlim(-2, 3) +  
 geom\_abline(intercept = 0,  
 slope = 1,  
 color = "red",  
 size = 2)

## Warning: Removed 2 rows containing missing values (geom\_point).



# Linear Regression | removed correlations

# Removed >.9 correlations  
correlation\_matrix = cor(train\_transformed)  
correlated\_columns = findCorrelation(correlation\_matrix, cutoff = 0.9)  
correlated\_columns = sort(correlated\_columns)  
train\_transformed\_no\_correlated = train\_transformed[,-c(correlated\_columns)]  
label\_index\_noCorr <- which(colnames(train\_transformed\_no\_correlated) == "critical\_temp")  
  
fit\_lm\_no\_correlated <- train(critical\_temp ~ ., data = train\_transformed\_no\_correlated, method = "lm", trControl = train\_control\_cv)  
print(fit\_lm\_no\_correlated)

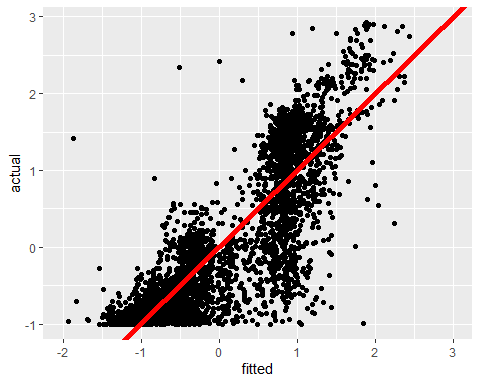
## Linear Regression   
##   
## 17011 samples  
## 123 predictor  
##   
## No pre-processing  
## Resampling: Cross-Validated (10 fold)   
## Summary of sample sizes: 15310, 15310, 15310, 15310, 15309, 15311, ...   
## Resampling results:  
##   
## RMSE Rsquared MAE   
## 0.591452 0.6984312 0.3787938  
##   
## Tuning parameter 'intercept' was held constant at a value of TRUE

test\_lm\_noCorr <- predict(fit\_lm\_no\_correlated, newdata = test\_transformed[, -c(correlated\_columns)])  
round(postResample(pred = test\_lm\_noCorr, obs = test\_transformed$critical\_temp), 3)

## RMSE Rsquared MAE   
## 0.937 0.399 0.393

fitted\_actual\_lm\_noCorr <- data.frame(fitted = test\_lm\_noCorr, actual = test\_transformed$critical\_temp)  
  
ggplot(fitted\_actual\_lm\_noCorr,   
 aes(x = fitted,  
 y = actual)) +  
 geom\_point() +  
 xlim(-2, 3) +  
 geom\_abline(intercept = 0,  
 slope = 1,  
 color = "red",  
 size = 2)

## Warning: Removed 2 rows containing missing values (geom\_point).



# Random Forest

library(randomForest)

## Warning: package 'randomForest' was built under R version 4.0.5

## randomForest 4.6-14

## Type rfNews() to see new features/changes/bug fixes.

##   
## Attaching package: 'randomForest'

## The following object is masked from 'package:ggplot2':  
##   
## margin

## The following object is masked from 'package:dplyr':  
##   
## combine

fit\_rf <- randomForest(x = train\_transformed[, -label\_index], y = train\_transformed[, label\_index], ntree = 150, do.trace = T)

## | Out-of-bag |  
## Tree | MSE %Var(y) |  
## 1 | 0.1359 13.59 |  
## 2 | 0.136 13.60 |  
## 3 | 0.1244 12.44 |  
## 4 | 0.1168 11.68 |  
## 5 | 0.1117 11.17 |  
## 6 | 0.1086 10.86 |  
## 7 | 0.104 10.40 |  
## 8 | 0.09868 9.87 |  
## 9 | 0.09559 9.56 |  
## 10 | 0.0944 9.44 |  
## 11 | 0.09213 9.21 |  
## 12 | 0.0905 9.05 |  
## 13 | 0.08912 8.91 |  
## 14 | 0.08748 8.75 |  
## 15 | 0.08611 8.61 |  
## 16 | 0.08482 8.48 |  
## 17 | 0.08395 8.40 |  
## 18 | 0.08289 8.29 |  
## 19 | 0.08208 8.21 |  
## 20 | 0.08188 8.19 |  
## 21 | 0.08161 8.16 |  
## 22 | 0.081 8.10 |  
## 23 | 0.08048 8.05 |  
## 24 | 0.08032 8.03 |  
## 25 | 0.0801 8.01 |  
## 26 | 0.07926 7.93 |  
## 27 | 0.07867 7.87 |  
## 28 | 0.07864 7.86 |  
## 29 | 0.07831 7.83 |  
## 30 | 0.07801 7.80 |  
## 31 | 0.07758 7.76 |  
## 32 | 0.07732 7.73 |  
## 33 | 0.07711 7.71 |  
## 34 | 0.07677 7.68 |  
## 35 | 0.07641 7.64 |  
## 36 | 0.07612 7.61 |  
## 37 | 0.0758 7.58 |  
## 38 | 0.0757 7.57 |  
## 39 | 0.07564 7.56 |  
## 40 | 0.07555 7.56 |  
## 41 | 0.07543 7.54 |  
## 42 | 0.07518 7.52 |  
## 43 | 0.075 7.50 |  
## 44 | 0.07508 7.51 |  
## 45 | 0.07502 7.50 |  
## 46 | 0.07494 7.49 |  
## 47 | 0.07493 7.49 |  
## 48 | 0.07485 7.49 |  
## 49 | 0.07477 7.48 |  
## 50 | 0.07472 7.47 |  
## 51 | 0.07469 7.47 |  
## 52 | 0.07453 7.45 |  
## 53 | 0.0743 7.43 |  
## 54 | 0.07425 7.43 |  
## 55 | 0.07413 7.41 |  
## 56 | 0.07402 7.40 |  
## 57 | 0.07409 7.41 |  
## 58 | 0.07397 7.40 |  
## 59 | 0.07396 7.40 |  
## 60 | 0.07395 7.40 |  
## 61 | 0.07385 7.39 |  
## 62 | 0.07372 7.37 |  
## 63 | 0.07374 7.37 |  
## 64 | 0.07371 7.37 |  
## 65 | 0.07365 7.37 |  
## 66 | 0.0736 7.36 |  
## 67 | 0.07347 7.35 |  
## 68 | 0.07341 7.34 |  
## 69 | 0.07323 7.32 |  
## 70 | 0.07324 7.32 |  
## 71 | 0.07316 7.32 |  
## 72 | 0.07316 7.32 |  
## 73 | 0.07319 7.32 |  
## 74 | 0.07314 7.31 |  
## 75 | 0.07314 7.31 |  
## 76 | 0.07305 7.31 |  
## 77 | 0.07294 7.29 |  
## 78 | 0.07288 7.29 |  
## 79 | 0.07278 7.28 |  
## 80 | 0.07284 7.28 |  
## 81 | 0.07278 7.28 |  
## 82 | 0.07277 7.28 |  
## 83 | 0.0728 7.28 |  
## 84 | 0.07284 7.28 |  
## 85 | 0.0728 7.28 |  
## 86 | 0.07268 7.27 |  
## 87 | 0.07271 7.27 |  
## 88 | 0.07264 7.26 |  
## 89 | 0.07261 7.26 |  
## 90 | 0.0726 7.26 |  
## 91 | 0.07261 7.26 |  
## 92 | 0.07251 7.25 |  
## 93 | 0.07254 7.25 |  
## 94 | 0.07254 7.25 |  
## 95 | 0.07256 7.26 |  
## 96 | 0.07249 7.25 |  
## 97 | 0.07246 7.25 |  
## 98 | 0.07243 7.24 |  
## 99 | 0.0724 7.24 |  
## 100 | 0.07238 7.24 |  
## 101 | 0.07236 7.24 |  
## 102 | 0.07236 7.24 |  
## 103 | 0.07234 7.23 |  
## 104 | 0.07238 7.24 |  
## 105 | 0.07239 7.24 |  
## 106 | 0.07241 7.24 |  
## 107 | 0.07238 7.24 |  
## 108 | 0.07227 7.23 |  
## 109 | 0.07222 7.22 |  
## 110 | 0.07221 7.22 |  
## 111 | 0.0722 7.22 |  
## 112 | 0.07213 7.21 |  
## 113 | 0.07209 7.21 |  
## 114 | 0.07214 7.21 |  
## 115 | 0.0721 7.21 |  
## 116 | 0.07211 7.21 |  
## 117 | 0.07211 7.21 |  
## 118 | 0.07206 7.21 |  
## 119 | 0.07208 7.21 |  
## 120 | 0.07212 7.21 |  
## 121 | 0.07209 7.21 |  
## 122 | 0.07207 7.21 |  
## 123 | 0.0721 7.21 |  
## 124 | 0.07208 7.21 |  
## 125 | 0.07209 7.21 |  
## 126 | 0.07209 7.21 |  
## 127 | 0.07208 7.21 |  
## 128 | 0.07209 7.21 |  
## 129 | 0.072 7.20 |  
## 130 | 0.07196 7.20 |  
## 131 | 0.07195 7.20 |  
## 132 | 0.07192 7.19 |  
## 133 | 0.07192 7.19 |  
## 134 | 0.0719 7.19 |  
## 135 | 0.07192 7.19 |  
## 136 | 0.07194 7.19 |  
## 137 | 0.07191 7.19 |  
## 138 | 0.07193 7.19 |  
## 139 | 0.07196 7.20 |  
## 140 | 0.07192 7.19 |  
## 141 | 0.0719 7.19 |  
## 142 | 0.07187 7.19 |  
## 143 | 0.07186 7.19 |  
## 144 | 0.07182 7.18 |  
## 145 | 0.07184 7.18 |  
## 146 | 0.07177 7.18 |  
## 147 | 0.07179 7.18 |  
## 148 | 0.0718 7.18 |  
## 149 | 0.07181 7.18 |  
## 150 | 0.07184 7.18 |

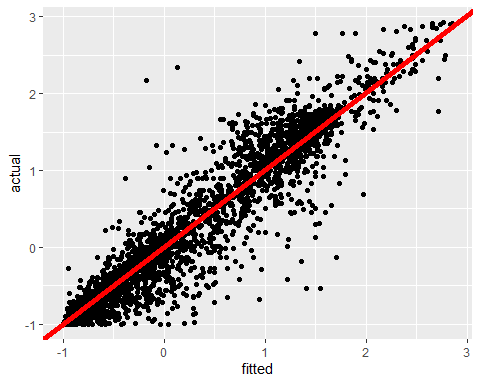
print(fit\_rf)

##   
## Call:  
## randomForest(x = train\_transformed[, -label\_index], y = train\_transformed[, label\_index], ntree = 150, do.trace = T)   
## Type of random forest: regression  
## Number of trees: 150  
## No. of variables tried at each split: 52  
##   
## Mean of squared residuals: 0.07184329  
## % Var explained: 92.82

test\_rf = predict(fit\_rf, newdata = test\_transformed[, -label\_index])  
round(postResample(pred = test\_rf, obs = test\_transformed$critical\_temp), 3)

## RMSE Rsquared MAE   
## 0.263 0.929 0.151

fitted\_actual\_rf <- data.frame(fitted = test\_rf, actual = test\_transformed$critical\_temp)  
  
ggplot(fitted\_actual\_rf,   
 aes(x = fitted,  
 y = actual)) +  
 geom\_point() +  
 geom\_abline(intercept = 0,  
 slope = 1,  
 color = "red",  
 size = 2)



# Random Forest | removed correlations

library(randomForest)  
fit\_rf\_noCorr <- randomForest(x = train\_transformed\_no\_correlated[, -label\_index\_noCorr], y = train\_transformed\_no\_correlated[, label\_index\_noCorr], ntree = 150, do.trace = T)

## | Out-of-bag |  
## Tree | MSE %Var(y) |  
## 1 | 0.1385 13.85 |  
## 2 | 0.1365 13.65 |  
## 3 | 0.1239 12.39 |  
## 4 | 0.1155 11.55 |  
## 5 | 0.1102 11.02 |  
## 6 | 0.1044 10.44 |  
## 7 | 0.1018 10.18 |  
## 8 | 0.09889 9.89 |  
## 9 | 0.09515 9.52 |  
## 10 | 0.09295 9.30 |  
## 11 | 0.09059 9.06 |  
## 12 | 0.0882 8.82 |  
## 13 | 0.08696 8.70 |  
## 14 | 0.08573 8.57 |  
## 15 | 0.08395 8.40 |  
## 16 | 0.08335 8.34 |  
## 17 | 0.08236 8.24 |  
## 18 | 0.0818 8.18 |  
## 19 | 0.08102 8.10 |  
## 20 | 0.08062 8.06 |  
## 21 | 0.08002 8.00 |  
## 22 | 0.07973 7.97 |  
## 23 | 0.07914 7.91 |  
## 24 | 0.07885 7.89 |  
## 25 | 0.07856 7.86 |  
## 26 | 0.07808 7.81 |  
## 27 | 0.07756 7.76 |  
## 28 | 0.07741 7.74 |  
## 29 | 0.07712 7.71 |  
## 30 | 0.07693 7.69 |  
## 31 | 0.07688 7.69 |  
## 32 | 0.07677 7.68 |  
## 33 | 0.07676 7.68 |  
## 34 | 0.07651 7.65 |  
## 35 | 0.0763 7.63 |  
## 36 | 0.07634 7.63 |  
## 37 | 0.07624 7.62 |  
## 38 | 0.07602 7.60 |  
## 39 | 0.07577 7.58 |  
## 40 | 0.07554 7.55 |  
## 41 | 0.07549 7.55 |  
## 42 | 0.07554 7.55 |  
## 43 | 0.07537 7.54 |  
## 44 | 0.07503 7.50 |  
## 45 | 0.07493 7.49 |  
## 46 | 0.07497 7.50 |  
## 47 | 0.07482 7.48 |  
## 48 | 0.07466 7.47 |  
## 49 | 0.07451 7.45 |  
## 50 | 0.0744 7.44 |  
## 51 | 0.07428 7.43 |  
## 52 | 0.07421 7.42 |  
## 53 | 0.07405 7.41 |  
## 54 | 0.07401 7.40 |  
## 55 | 0.07405 7.41 |  
## 56 | 0.07389 7.39 |  
## 57 | 0.07389 7.39 |  
## 58 | 0.07377 7.38 |  
## 59 | 0.07367 7.37 |  
## 60 | 0.07358 7.36 |  
## 61 | 0.07359 7.36 |  
## 62 | 0.07354 7.35 |  
## 63 | 0.07351 7.35 |  
## 64 | 0.07359 7.36 |  
## 65 | 0.07345 7.35 |  
## 66 | 0.07353 7.35 |  
## 67 | 0.07353 7.35 |  
## 68 | 0.07345 7.35 |  
## 69 | 0.07341 7.34 |  
## 70 | 0.07336 7.34 |  
## 71 | 0.07333 7.33 |  
## 72 | 0.07336 7.34 |  
## 73 | 0.07326 7.33 |  
## 74 | 0.07328 7.33 |  
## 75 | 0.07324 7.32 |  
## 76 | 0.0732 7.32 |  
## 77 | 0.0732 7.32 |  
## 78 | 0.07312 7.31 |  
## 79 | 0.07314 7.31 |  
## 80 | 0.07302 7.30 |  
## 81 | 0.07292 7.29 |  
## 82 | 0.07288 7.29 |  
## 83 | 0.07281 7.28 |  
## 84 | 0.0728 7.28 |  
## 85 | 0.0727 7.27 |  
## 86 | 0.07267 7.27 |  
## 87 | 0.07264 7.26 |  
## 88 | 0.0726 7.26 |  
## 89 | 0.07251 7.25 |  
## 90 | 0.07257 7.26 |  
## 91 | 0.07256 7.26 |  
## 92 | 0.07249 7.25 |  
## 93 | 0.07253 7.25 |  
## 94 | 0.07252 7.25 |  
## 95 | 0.07255 7.26 |  
## 96 | 0.0725 7.25 |  
## 97 | 0.07251 7.25 |  
## 98 | 0.07249 7.25 |  
## 99 | 0.0725 7.25 |  
## 100 | 0.07249 7.25 |  
## 101 | 0.07243 7.24 |  
## 102 | 0.07244 7.24 |  
## 103 | 0.07243 7.24 |  
## 104 | 0.07242 7.24 |  
## 105 | 0.07233 7.23 |  
## 106 | 0.07227 7.23 |  
## 107 | 0.0722 7.22 |  
## 108 | 0.07218 7.22 |  
## 109 | 0.07213 7.21 |  
## 110 | 0.07214 7.21 |  
## 111 | 0.07216 7.22 |  
## 112 | 0.0722 7.22 |  
## 113 | 0.07222 7.22 |  
## 114 | 0.07221 7.22 |  
## 115 | 0.07215 7.22 |  
## 116 | 0.0722 7.22 |  
## 117 | 0.07223 7.22 |  
## 118 | 0.07219 7.22 |  
## 119 | 0.07222 7.22 |  
## 120 | 0.07221 7.22 |  
## 121 | 0.07223 7.22 |  
## 122 | 0.07221 7.22 |  
## 123 | 0.07219 7.22 |  
## 124 | 0.07217 7.22 |  
## 125 | 0.07219 7.22 |  
## 126 | 0.07218 7.22 |  
## 127 | 0.07221 7.22 |  
## 128 | 0.07219 7.22 |  
## 129 | 0.07217 7.22 |  
## 130 | 0.07221 7.22 |  
## 131 | 0.0722 7.22 |  
## 132 | 0.07218 7.22 |  
## 133 | 0.07216 7.22 |  
## 134 | 0.07216 7.22 |  
## 135 | 0.07211 7.21 |  
## 136 | 0.07212 7.21 |  
## 137 | 0.07215 7.22 |  
## 138 | 0.07214 7.21 |  
## 139 | 0.07211 7.21 |  
## 140 | 0.07211 7.21 |  
## 141 | 0.07209 7.21 |  
## 142 | 0.07206 7.21 |  
## 143 | 0.07207 7.21 |  
## 144 | 0.07206 7.21 |  
## 145 | 0.07204 7.20 |  
## 146 | 0.07201 7.20 |  
## 147 | 0.07199 7.20 |  
## 148 | 0.07195 7.20 |  
## 149 | 0.07193 7.19 |  
## 150 | 0.07197 7.20 |

print(fit\_rf\_noCorr)

##   
## Call:  
## randomForest(x = train\_transformed\_no\_correlated[, -label\_index\_noCorr], y = train\_transformed\_no\_correlated[, label\_index\_noCorr], ntree = 150, do.trace = T)   
## Type of random forest: regression  
## Number of trees: 150  
## No. of variables tried at each split: 41  
##   
## Mean of squared residuals: 0.07196968  
## % Var explained: 92.8

test\_rf\_noCorr = predict(fit\_rf\_noCorr, newdata = test\_transformed[, -c(correlated\_columns,label\_index)])  
round(postResample(pred = test\_rf\_noCorr, obs = test\_transformed$critical\_temp), 3)

## RMSE Rsquared MAE   
## 0.263 0.929 0.151

fitted\_actual\_rf\_noCorr <- data.frame(fitted = test\_rf\_noCorr, actual = test\_transformed$critical\_temp)  
  
ggplot(fitted\_actual\_rf\_noCorr,   
 aes(x = fitted,  
 y = actual)) +  
 geom\_point() +  
 geom\_abline(intercept = 0,  
 slope = 1,  
 color = "red",  
 size = 2)

