ames iowa kaggle home price modeling

preeti swaminathan & patrick mcdevitt

14 avril 2017

House Prices - à la Kaggle - Homework 13

Analysis Question 2

```
    setwd(home_dir)
    setwd(data_dir)

    homes <- read.csv("train.csv", stringsAsFactors = FALSE)
    setwd(home_dir)

    names(homes) <- tolower(names(homes))

    for (i in 2:(length(homes)))
    {
        if (class(homes[,i]) == "character")
        {
            homes[,i] <- factor (homes[,i])
        }
    }
```

```
# ...   -=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-
# ...   remove outliers ... more than 5 sigma from mean value
# ...   -=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-

    lst <- length(homes) - 1    # sale price is (currently) last column

    for (i in 2 : lst)
    {
        if(class(homes[,i]) == "integer" || class(homes[,i]) == "numeric")
        {
            homes[,i][which(scale(homes[,i]) > 5)] <- NA
            homes[,i][which(scale(homes[,i]) < -5)] <- NA
        }
    }
```

```
# ...   -=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-
# ...   create a few new columns
# ...   -=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-

    dates <- paste(homes$yrsold, sprintf("%02d", homes$mosold), "01")
    homes$sale_date <- as.Date(dates, "%Y %m %d")

# ...   -=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-
# ...   scale each column independently
# ...   -=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-

#   for (i in 2 : length(homes))
#   {
#       if(class(homes[,i]) == "integer" || class(homes[,i]) == "numeric")
#       {
#           homes[,i] <- scale(homes[,i])
#       }
#   }
```

```
# ...   -=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-
# ...   make some plots for numberic variables... linear, log_x, log_y, log_xy ...
# ...   -=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-

    pdf ("homes_train_plots.pdf", width = 10, height = 7)

    par (mfrow = c (2, 3))
    for (i in 2:(length(homes)))
    {
        if(class(homes[,i]) == "integer" || class(homes[,i]) == "numeric" || class(homes[,i]) == "matri
        {
            plot (homes[,i], main = (names(homes[i])))
            hist(homes[,i])
            plot(log(homes$saleprice)  ~ homes[,i])
        }
    }

    par (mfrow = c (2, 2))
    for (i in 2:(length(homes)))
    {
        if(class(homes[,i]) == "factor")
        {
            plot_title <- names(homes[i])

            p <- ggplot(homes, aes(x = homes[,i], fill = homes[,i])) + geom_bar() + labs(title = plot_ti
            print(p)

            p <- ggplot(homes, aes(x = homes[,i], y = log(saleprice), fill = homes[,i])) + geom_boxplot
            print(p)
        }
    }

            plot(log(homes$saleprice) ~ homes$sale_date)

    dev.off()
```

```
## pdf
##   2
```

```
    for (i in 2:(length(homes)))
    {
        if(class(homes[,i]) == "integer" || class(homes[,i]) == "numeric" || class(homes[,i]) == "matri
        {
            fit <- lm(log(homes$saleprice) ~ homes[,i])

            print(sprintf(" ... %3d : %20s | %10s | r^2 = %8.3f | p-value = %12.4e",
                          i, names(homes[i]), class(homes[,i]), summary(fit)$r.squared, summary(fit)$coe
        }
```

```r
    }
```

```
## [1] " ...   2 :            mssubclass |    integer | r^2 =    0.005 | p-value =   4.6924e-03"
## [1] " ...   4 :           lotfrontage |    integer | r^2 =    0.146 | p-value =   4.7860e-43"
## [1] " ...   5 :               lotarea |    integer | r^2 =    0.135 | p-value =   1.3152e-47"
## [1] " ...  18 :           overallqual |    integer | r^2 =    0.668 | p-value =   0.0000e+00"
## [1] " ...  19 :           overallcond |    integer | r^2 =    0.001 | p-value =   1.5913e-01"
## [1] " ...  20 :             yearbuilt |    integer | r^2 =    0.344 | p-value =  1.1036e-135"
## [1] " ...  21 :          yearremodadd |    integer | r^2 =    0.320 | p-value =  3.2115e-124"
## [1] " ...  27 :            masvnrarea |    integer | r^2 =    0.183 | p-value =   2.1180e-65"
## [1] " ...  35 :            bsmtfinsf1 |    integer | r^2 =    0.153 | p-value =   2.1664e-54"
## [1] " ...  37 :            bsmtfinsf2 |    integer | r^2 =    0.002 | p-value =   1.1873e-01"
## [1] " ...  38 :             bsmtunfsf |    integer | r^2 =    0.049 | p-value =   9.3185e-18"
## [1] " ...  39 :           totalbsmtsf |    integer | r^2 =    0.413 | p-value =  6.2898e-171"
## [1] " ...  44 :             x1stflrsf |    integer | r^2 =    0.383 | p-value =  5.3102e-155"
## [1] " ...  45 :             x2ndflrsf |    integer | r^2 =    0.102 | p-value =   5.8669e-36"
## [1] " ...  46 :          lowqualfinsf |    integer | r^2 =    0.004 | p-value =   1.1152e-02"
## [1] " ...  47 :             grlivarea |    integer | r^2 =    0.517 | p-value =  7.3321e-232"
## [1] " ...  48 :          bsmtfullbath |    integer | r^2 =    0.056 | p-value =   5.7917e-20"
## [1] " ...  49 :          bsmthalfbath |    integer | r^2 =    0.000 | p-value =   8.8755e-01"
## [1] " ...  50 :              fullbath |    integer | r^2 =    0.354 | p-value =  2.1190e-140"
## [1] " ...  51 :              halfbath |    integer | r^2 =    0.099 | p-value =   9.1331e-35"
## [1] " ...  52 :          bedroomabvgr |    integer | r^2 =    0.044 | p-value =   5.3387e-16"
## [1] " ...  53 :          kitchenabvgr |    integer | r^2 =    0.021 | p-value =   2.0002e-08"
## [1] " ...  55 :          totrmsabvgrd |    integer | r^2 =    0.286 | p-value =  1.2928e-108"
## [1] " ...  57 :            fireplaces |    integer | r^2 =    0.240 | p-value =   8.4213e-89"
## [1] " ...  60 :           garageyrblt |    integer | r^2 =    0.293 | p-value =  1.0597e-105"
## [1] " ...  62 :            garagecars |    integer | r^2 =    0.463 | p-value =  3.0938e-199"
## [1] " ...  63 :            garagearea |    integer | r^2 =    0.424 | p-value =  1.1063e-176"
## [1] " ...  67 :            wooddecksf |    integer | r^2 =    0.114 | p-value =   2.4587e-40"
## [1] " ...  68 :           openporchsf |    integer | r^2 =    0.126 | p-value =   1.3467e-44"
## [1] " ...  69 :         enclosedporch |    integer | r^2 =    0.027 | p-value =   1.9537e-10"
## [1] " ...  70 :            x3ssnporch |    integer | r^2 =    0.000 | p-value =   6.2057e-01"
## [1] " ...  71 :           screenporch |    integer | r^2 =    0.009 | p-value =   2.8727e-04"
## [1] " ...  72 :              poolarea |    integer | r^2 =    0.000 | p-value =           NA"
## [1] " ...  76 :               miscval |    integer | r^2 =    0.000 | p-value =   5.4069e-01"
## [1] " ...  77 :                mosold |    integer | r^2 =    0.003 | p-value =   2.8489e-02"
## [1] " ...  78 :                yrsold |    integer | r^2 =    0.001 | p-value =   1.5471e-01"
## [1] " ...  81 :             saleprice |    integer | r^2 =    0.899 | p-value =   0.0000e+00"
```

```r
# ...   -=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-
# ...   Impute NAs to functional value
# ...
# ...   --> for numerical variables - impute to mean value in column
# ...   --> for factor variables - create new factor "None"
# ...   -=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-

    for (i in 1 : (length(homes)))
    {
        if(class(homes[,i]) == "integer" || class(homes[,i]) == "numeric" || class(homes[,i]) == "matri
        {
            homes[,i][is.na (homes[,i])] <- mean(homes[,i], na.rm = TRUE)
        }
    }
    for (i in 1 : (length(homes)))
    {
        if(class(homes[,i]) == "factor")
        {
            levels <- levels(homes[,i])
            levels[length(levels) + 1] <- "None"
            homes[,i] <- factor(homes[,i], levels = levels)
            homes[,i][is.na (homes[,i])] <- "None"
        }
    }
```

```
## Warning in `levels<-`(`*tmp*`, value = if (nl == nL) as.character(labels)
## else paste0(labels, : duplicated levels in factors are deprecated
```

```
    for (i in 2:(length(homes)))
    {
        if(class(homes[,i]) == "integer" || class(homes[,i]) == "numeric" || class(homes[,i]) == "matri
        {
            fit <- lm(log(homes$saleprice) ~ homes[,i])

            print(sprintf(" ... %3d : %20s | %10s | r^2 = %8.3f | p-value = %12.4e",
                          i, names(homes[i]), class(homes[,i]), summary(fit)$r.squared, summary(fit)$coe
        }
    }
```

```
## [1] " ...   2 :           mssubclass |    numeric | r^2 =    0.005 | p-value =   4.6924e-03"
## [1] " ...   4 :          lotfrontage |    numeric | r^2 =    0.130 | p-value =   3.3560e-46"
## [1] " ...   5 :              lotarea |    numeric | r^2 =    0.134 | p-value =   1.9014e-47"
## [1] " ...  18 :          overallqual |    numeric | r^2 =    0.668 | p-value =   0.0000e+00"
## [1] " ...  19 :          overallcond |    numeric | r^2 =    0.001 | p-value =   1.5913e-01"
## [1] " ...  20 :            yearbuilt |    numeric | r^2 =    0.344 | p-value =  1.1036e-135"
## [1] " ...  21 :         yearremodadd |    numeric | r^2 =    0.320 | p-value =  3.2115e-124"
## [1] " ...  27 :           masvnrarea |    numeric | r^2 =    0.178 | p-value =   4.9217e-64"
## [1] " ...  35 :           bsmtfinsf1 |    numeric | r^2 =    0.153 | p-value =   1.9954e-54"
## [1] " ...  37 :           bsmtfinsf2 |    numeric | r^2 =    0.002 | p-value =   1.1824e-01"
## [1] " ...  38 :            bsmtunfsf |    numeric | r^2 =    0.049 | p-value =   9.3185e-18"
## [1] " ...  39 :          totalbsmtsf |    numeric | r^2 =    0.413 | p-value =  4.8341e-171"
## [1] " ...  44 :            x1stflrsf |    numeric | r^2 =    0.382 | p-value =  1.5129e-154"
## [1] " ...  45 :            x2ndflrsf |    numeric | r^2 =    0.102 | p-value =   5.8669e-36"
## [1] " ...  46 :         lowqualfinsf |    numeric | r^2 =    0.004 | p-value =   1.1233e-02"
## [1] " ...  47 :            grlivarea |    numeric | r^2 =    0.507 | p-value =  5.9322e-226"
## [1] " ...  48 :         bsmtfullbath |    numeric | r^2 =    0.056 | p-value =   5.7917e-20"
## [1] " ...  49 :         bsmthalfbath |    numeric | r^2 =    0.000 | p-value =   8.8749e-01"
## [1] " ...  50 :             fullbath |    numeric | r^2 =    0.354 | p-value =  2.1190e-140"
## [1] " ...  51 :             halfbath |    numeric | r^2 =    0.099 | p-value =   9.1331e-35"
## [1] " ...  52 :         bedroomabvgr |    numeric | r^2 =    0.044 | p-value =   5.2433e-16"
## [1] " ...  53 :         kitchenabvgr |    numeric | r^2 =    0.021 | p-value =   1.9939e-08"
## [1] " ...  55 :         totrmsabvgrd |    numeric | r^2 =    0.286 | p-value =  1.2928e-108"
## [1] " ...  57 :           fireplaces |    numeric | r^2 =    0.240 | p-value =   8.4213e-89"
## [1] " ...  60 :          garageyrblt |    numeric | r^2 =    0.250 | p-value =   2.2375e-93"
## [1] " ...  62 :           garagecars |    numeric | r^2 =    0.463 | p-value =  3.0938e-199"
## [1] " ...  63 :           garagearea |    numeric | r^2 =    0.424 | p-value =  1.1063e-176"
## [1] " ...  67 :           wooddecksf |    numeric | r^2 =    0.114 | p-value =   2.7488e-40"
## [1] " ...  68 :          openporchsf |    numeric | r^2 =    0.125 | p-value =   3.9027e-44"
## [1] " ...  69 :         enclosedporch |    numeric | r^2 =    0.027 | p-value =   1.9335e-10"
## [1] " ...  70 :            x3ssnporch |    numeric | r^2 =    0.000 | p-value =   6.2049e-01"
## [1] " ...  71 :           screenporch |    numeric | r^2 =    0.009 | p-value =   2.9494e-04"
## [1] " ...  72 :              poolarea |    numeric | r^2 =    0.000 | p-value =           NA"
## [1] " ...  76 :               miscval |    numeric | r^2 =    0.000 | p-value =   5.4137e-01"
## [1] " ...  77 :                mosold |    numeric | r^2 =    0.003 | p-value =   2.8489e-02"
## [1] " ...  78 :                yrsold |    numeric | r^2 =    0.001 | p-value =   1.5471e-01"
## [1] " ...  81 :             saleprice |    numeric | r^2 =    0.899 | p-value =   0.0000e+00"
```

```
# ...   -=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-
# ...   Columns to remove - based on visual inspection
# ...   -=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-

# ...   save top 20 (based on r^2) for trial evaluation in SAS

    homes$log_lotfrontage <- log(homes$lotfrontage)
    homes$log_lotarea <- log(homes$lotarea)
    homes$log_grlivarea <- log(homes$grlivarea)
    homes$log_saleprice <- log(homes$saleprice)

    homes_sas_keep <- subset(homes,
            select = c(
                log_saleprice,
            bsmtfinsf1,
            bsmtfintype1,
            bsmtfullbath,
            bsmtqual,
            centralair,
            electrical,
            exterior1st,
            exterior2nd,
            exterqual,
            fireplacequ,
            fireplaces,
            foundation,
            fullbath,
            garagearea,
            garagecars,
            garagefinish,
            garagetype,
            grlivarea,
            halfbath,
            heatingqc,
            housestyle,
            kitchenqual,
            log_grlivarea,
            log_saleprice,
            log_lotarea,
            log_lotfrontage,
            lotshape,
            masvnrtype,
            mszoning,
            neighborhood,
            overallcond,
            overallqual,
            saletype,
            totalbsmtsf,
            totrmsabvgrd,
            x1stflrsf,
            x2ndflrsf,
            yearbuilt,
            yearremodadd))

# ...   -=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-
# ... from the keep list , these are the factors :
# ...   bsmtfintype1, bsmtqual, centralair, electrical, exterior1st, exterior2nd, exterqual,
# ...   fireplacequ, foundation, garagefinish, garagetype, heatingqc, housestyle, kitchenqual,
# ...   lotshape, masvnrtype, mszoning, neighborhood, saletype,
# ...   -=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-


# ...   -=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-
# ...   store reference data frame as base data set
# ...   -=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-

        homes_subset_base <- homes_sas_keep
```

```
# ...   -=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-
# ...   save data frame for SAS input file
# ...   -=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-

    sas_dir <- "~/sas/SASUniversityEdition/myfolders/"
    setwd(sas_dir)
    write.csv (homes_sas_keep, file = "training_set_cleaned.csv", row.names = FALSE)

    setwd(home_dir)
    setwd(data_dir)
    write.csv (homes_sas_keep, file = "training_set_cleaned.csv", row.names = FALSE)
    setwd(home_dir)
```