Extracting basic info from a data table- red wine data

HDS, Mike Kozlowski

9/5/2020

Updated 1/31/2023

Extracting Basic Information from a Data Table

We'll like at the red wine data set again.

Remember, even though the file type says it should be CSV, the separator here is a semicolon

I tend to immediately use summary() and str() on data I haven't seen before (or haven't seen lately)

file.choose() is used in the function call to read.table, file.choose() will open the file browser so you can locate the winequality-red file

```
myData<-read.table("C:\\Users\\Mike\\Documents\\DAT511\\2-1 class\\winequality-red1.csv", hea
der=TRUE, sep=";")
summary(myData)
```

```
##  fixed.acidity   volatile.acidity  citric.acid    residual.sugar
##  Min.   : 4.60   Min.   :0.1200   Min.   :0.000   Min.   : 0.900
##  1st Qu.: 7.10   1st Qu.:0.3900   1st Qu.:0.090   1st Qu.: 1.900
##  Median : 7.90   Median :0.5200   Median :0.260   Median : 2.200
##  Mean   : 8.32   Mean   :0.5278   Mean   :0.271   Mean   : 2.539
##  3rd Qu.: 9.20   3rd Qu.:0.6400   3rd Qu.:0.420   3rd Qu.: 2.600
##  Max.   :15.90   Max.   :1.5800   Max.   :1.000   Max.   :15.500
##    chlorides       free.sulfur.dioxide total.sulfur.dioxide    density
##  Min.   :0.01200   Min.   : 1.00       Min.   :  6.00       Min.   :0.9901
##  1st Qu.:0.07000   1st Qu.: 7.00       1st Qu.: 22.00       1st Qu.:0.9956
##  Median :0.07900   Median :14.00       Median : 38.00       Median :0.9968
##  Mean   :0.08747   Mean   :15.87       Mean   : 46.47       Mean   :0.9967
##  3rd Qu.:0.09000   3rd Qu.:21.00       3rd Qu.: 62.00       3rd Qu.:0.9978
##  Max.   :0.61100   Max.   :72.00       Max.   :289.00       Max.   :1.0037
##        pH          sulphates         alcohol         quality
##  Min.   :2.740   Min.   :0.3300   Min.   : 8.40   Min.   :3.000
##  1st Qu.:3.210   1st Qu.:0.5500   1st Qu.: 9.50   1st Qu.:5.000
##  Median :3.310   Median :0.6200   Median :10.20   Median :6.000
##  Mean   :3.311   Mean   :0.6581   Mean   :10.42   Mean   :5.636
##  3rd Qu.:3.400   3rd Qu.:0.7300   3rd Qu.:11.10   3rd Qu.:6.000
##  Max.   :4.010   Max.   :2.0000   Max.   :14.90   Max.   :8.000
```

```
str(myData)
```

```
## 'data.frame':    1599 obs. of  12 variables:
##  $ fixed.acidity       : num  7.4 7.8 7.8 11.2 7.4 7.4 7.9 7.3 7.8 7.5 ...
##  $ volatile.acidity    : num  0.7 0.88 0.76 0.28 0.7 0.66 0.6 0.65 0.58 0.5 ...
##  $ citric.acid         : num  0 0 0.04 0.56 0 0 0.06 0 0.02 0.36 ...
##  $ residual.sugar      : num  1.9 2.6 2.3 1.9 1.9 1.8 1.6 1.2 2 6.1 ...
##  $ chlorides           : num  0.076 0.098 0.092 0.075 0.076 0.075 0.069 0.065 0.073 0.071
...
##  $ free.sulfur.dioxide : num  11 25 15 17 11 13 15 15 9 17 ...
##  $ total.sulfur.dioxide: num  34 67 54 60 34 40 59 21 18 102 ...
##  $ density             : num  0.998 0.997 0.997 0.998 0.998 ...
##  $ pH                  : num  3.51 3.2 3.26 3.16 3.51 3.51 3.3 3.39 3.36 3.35 ...
##  $ sulphates           : num  0.56 0.68 0.65 0.58 0.56 0.56 0.46 0.47 0.57 0.8 ...
##  $ alcohol             : num  9.4 9.8 9.8 9.8 9.4 9.4 9.4 10 9.5 10.5 ...
##  $ quality             : int  5 5 5 6 5 5 5 7 7 5 ...
```

This data set doesn't have row names or identifiers for the wine, I"m just going to add an integer row ID

```
rownames(myData)<-1:1599
```

#Tables

Let's create a table of the mean fixed.acidity for each quality grade

Here, we need to use quality as a factor, but it's currently an integer, we will to convert it to a factor to use it as the sorting variable in tapply

Remember in tapply, it is the data, then the factor, then the function to use in the calculation

```
tapply(myData$fixed.acidity,as.factor(myData$quality),mean)
```

```
##        3        4        5        6        7        8
## 8.360000 7.779245 8.167254 8.347179 8.872362 8.566667
```

We can see here that there are 6 levels of wine quality in the data set, we could get this from the unique() function as well, which gives us the unique values of a variable present- this is most effective with strings and factors, also with dates

```
unique(myData$quality)
```

```
## [1] 5 6 7 4 8 3
```

```
length(unique(myData$quality))
```

```
## [1] 6
```

#Question/Action

Create code to make two tables that show the min and the max alcohol levels by quality

```
tapply(myData$alcohol,as.factor(myData$quality),max)
```

```
##    3    4    5    6    7    8
## 11.0 13.1 14.9 14.0 14.0 14.0
```

```
tapply(myData$alcohol,as.factor(myData$quality),min)
```

```
##   3   4   5   6   7   8
## 8.4 9.0 8.5 8.4 9.2 9.8
```

We can use tapply to extract more than one piece of information at a time from a subset of the data.

Here I defined a function that computed a number of summary statistics related to the data

```
my_extract<-function(x)
{
  a=c(mean(x),sd(x),max(x),min(x))
  names(a)<-c("Mean","SD","Max","Min")
  return(a)
}
tapply(myData$fixed.acidity,as.factor(myData$quality),my_extract)
```

```
## $`3`
##      Mean        SD       Max       Min
##  8.360000  1.770875 11.600000  6.700000
##
## $`4`
##      Mean        SD       Max       Min
##  7.779245  1.626624 12.500000  4.600000
##
## $`5`
##      Mean        SD       Max       Min
##  8.167254  1.563988 15.900000  5.000000
##
## $`6`
##      Mean        SD       Max       Min
##  8.347179  1.797849 14.300000  4.700000
##
## $`7`
##      Mean        SD       Max       Min
##  8.872362  1.992483 15.600000  4.900000
##
## $`8`
##      Mean        SD       Max       Min
##  8.566667  2.119656 12.600000  5.000000
```

Question/Action

Load up the 2020-2021 assessment role. read.csv is sometimes more tolerant of data issues than read.table

```
myData2 = read.csv("C:\\Users\\Mike\\Documents\\DAT511\\2-1 class\\2020-2021_Assessment_Roll.
csv")
```

Check the structure

How many different property class descriptions are there? 145 total unique property descriptions

```
unique(myData2$PROP.CLASS.DESCRIPTION)
```

```
##   [1] "RESIDENTIAL VACANT LAND"
##   [2] "APARTMENT"
##   [3] "ONE FAMILY DWELLING"
##   [4] "COM VAC W/IMP"
##   [5] "COMMERCIAL VACANT LAND"
##   [6] "TWO FAMILY DWELLING"
##   [7] "OFFICE BUILDING"
##   [8] "HOSPITALS"
##   [9] "GAS MEAS STATION"
##  [10] "OTHER STORAGE & WAREHOUSE FACILITIES"
##  [11] "RESIDENTIAL LAND WITH SMALL IMPROVEMENTS"
##  [12] "RELIGIOUS"
##  [13] "CELL TOWER"
##  [14] "AUTO DEALERS"
##  [15] "PARKING LOT"
##  [16] "THREE FAMILY DWELLING"
##  [17] "GOVERNMENTAL CENTERS"
##  [18] "NON-CEILING RAILROADS"
##  [19] "AUTO BODY AND TIRE SHOPS"
##  [20] "DOWNTOWN ROW TYPE (DETACHED)"
##  [21] "INDUSTRIAL VACANT LAND"
##  [22] "URBAN RENEWAL VACANT LAND"
##  [23] "RESTAURANTS"
##  [24] "TELEPHONE - SPECIAL FRANCHISE"
##  [25] "SCHOOL"
##  [26] "TELEPHONE"
##  [27] "MARINAS"
##  [28] "GAS OUTSIDE PLANT"
##  [29] "MANUFACTURING & PROCESSING"
##  [30] "CITY/TOWN/VILLAGE PUBLIC PARKS"
##  [31] "CEILING RAILROAD"
##  [32] "MULTIPLE RESIDENCES"
##  [33] "ELEC TRANS IMP"
##  [34] "FUNERAL HOMES"
##  [35] "RESIDENCE WITH COMMERCIAL USE"
##  [36] "ONE STORY SMALL STRUCTURE"
##  [37] "SERVICE AND GAS STATIONS"
##  [38] "AUTOMATIC CAR WASH"
##  [39] "ONE STORY SMALL STRUCTURE MULTI-OCCUPANT"
##  [40] "FAST FOOD FACILITY"
##  [41] "PROFESSIONAL BUILDING"
##  [42] "DOWNTOWN ROW TYPE (W/COMMON WALL)"
##  [43] "RECREATIONAL FACILITIES"
##  [44] "SOCIAL ORGANIZATIONS"
##  [45] "PARKING GARAGE"
##  [46] "AREA OR NEIGHBORHOOD SHOPPING CENTERS"
##  [47] "CONVERTED RESIDENCE"
##  [48] "BENEVOLENT AND MORAL ASSOCIATIONS"
##  [49] "SPECIAL SCHOOLS AND INSTITUTIONS"
##  [50] "LARGE RETAIL FOOD STORES"
##  [51] "BARS"
```

```
##  [52] "DINERS OR LUNCHEONETTES"
##  [53] "MISCELLANEOUS"
##  [54] "DRIVE-IN BANK BRANCH"
##  [55] "ALL OTHER HEALTH FACILITIES"
##  [56] "COMMUNITY SERVICES"
##  [57] "LIBRARY"
##  [58] "SMALL GARAGE"
##  [59] "SELF-SERVICE CAR WASH"
##  [60] "INDUSTRIAL VACANT LAND WITH IMPROVEMENTS"
##  [61] "INNS,LODGES,BOARDING AND ROOMING HOUSES"
##  [62] "ELEC DIST OUT"
##  [63] "HOME FOR AGED"
##  [64] "CULTURAL & RECREATIONAL FACILITIES"
##  [65] "ALL OTHER EDUCATIONAL FACILITIES"
##  [66] "COLLEGES AND UNIVERSITIES"
##  [67] "MOTOR VEHICLE"
##  [68] "CULTURAL FACILITIES"
##  [69] "GOVERNMENTAL BUILDINGS"
##  [70] "SMALL RETAIL"
##  [71] "MANUAL CAR WASH"
##  [72] "HOTEL"
##  [73] "MOTION PICTURE THEATER"
##  [74] "SWIMMING - INDOOR"
##  [75] "PARKING LOTS"
##  [76] "PLAYGROUNDS"
##  [77] "RADIO, TV, & MOTION PICTURES"
##  [78] "PARKS"
##  [79] "COMMUNICATIONS"
##  [80] "POLICE AND FIRE PROTECTION FACILITIES"
##  [81] "MEDIUM RETAIL"
##  [82] "ELEC PWR OTHR"
##  [83] "SKATING"
##  [84] "NIGHT CLUBS"
##  [85] "WELFARE"
##  [86] "TENNIS, ARCHERY, POOL, & BILLIARDS"
##  [87] "SNACK BARS,DRIVE-INS,ICE CREAM BARS"
##  [88] "DOG KENNELS, VETERINARY CLINICS"
##  [89] "BILLBOARDS"
##  [90] "GOVERNMENT HIGHWAY GARAGES"
##  [91] "ROADS,STREETS,HIGHWAYS & PARKWAYS"
##  [92] "CORRECTIONAL"
##  [93] "ARMY,MARINE & COAST GUARD INSTALLATIONS"
##  [94] "SINGLE FAMILY W/ APARTMENT"
##  [95] "MISCELLANEOUS SERVICES"
##  [96] "STADIUMS,ARENAS,ARMORY & FIELD HOUSES"
##  [97] "LARGE RETAIL OUTLET"
##  [98] "CEMETERIES"
##  [99] "BANK COMPLEX WITH OFFICE SPACE"
## [100] "LEGITIMATE THEATER"
## [101] "BOWLING"
## [102] "RADIO"
```

```
## [103] "TELEVISION OTHER THAN COMMUNITY ANTENNA"
## [104] "SKATING - OUTDOOR RINK"
## [105] "HEALTH"
## [106] "ANIMAL WELFARE"
## [107] "COLD STORAGE FACILITIES"
## [108] "AUDITORIUM,EXHIBITION OR EXPOSITION HALL"
## [109] "ATHLETIC FIELDS"
## [110] "LUMBER YARDS AND SAWMILLS"
## [111] "STANDARD BANK/SINGLE OCCUPANT"
## [112] "WATER SUPPLY"
## [113] "TRUCKING TERMINALS"
## [114] "ELEC - SUBSTATION"
## [115] "INDIAN RESERVATION"
## [116] "EDUCATION"
## [117] "GRAIN AND FEED ELEVATORS"
## [118] "WETLANDS - WILD OR CONSERVATION LANDS"
## [119] "LITE IND MANFTR"
## [120] "YMCA OR YWCA"
## [121] "SOLID WASTES"
## [122] "PUBLIC PARK"
## [123] "SEWAGE TREATMENT & WATER POLLUTION CNTRL"
## [124] "GAS TRANS IMPR"
## [125] "TRANSPORTATN"
## [126] "BOTTLED GAS, NATURAL GAS FACILITIES"
## [127] "GREENHOUSES"
## [128] "PIERS, WHARVES, DOCKS"
## [129] "GASOLINE,FUEL, & OIL STORAGE FACILITY"
## [130] "TELEVISION - SPECIAL FRANCHISE"
## [131] "ELECTRIC & GAS"
## [132] "PICKLE BALL COURT"
## [133] "OIL - FORCED"
## [134] "COUNTY OWNED PUBLIC PARKS"
## [135] "PETRO PIPELN"
## [136] "GOLF COURSES"
## [137] "WATER TREAT"
## [138] "MISC FRANCHS"
## [139] "SWIMMING - OUTDOOR POOL"
## [140] "WATER TRANS"
## [141] "COMMUNITY ANTENNA TELEVISION"
## [142] "HEALTH SPA"
## [143] "MOTOR VEHICLE SERVICES"
## [144] "DEALERSHIPS - SALES AND SERVICE"
## [145] "RIDING STABLES"
## [146] "BRIDGES,TUNNELS & SUBWAYS"
## [147] "STATE PARK"
## [148] "MOTEL"
```

How many distinct neighborhoods? 37 total unique neighborhoods

```
unique(myData2$NEIGHBORHOOD)
```

```
##  [1] "Fruit Belt"         "Masten Park"        "Allentown"
##  [4] "Elmwood Bidwell"    "Broadway Fillmore"  "MLK Park"
##  [7] "Elmwood Bryant"     "Delavan Grider"     "Central"
## [10] "Hamlin Park"        "West Side"          "Lower West Side"
## [13] "Genesee-Moselle"    "Schiller Park"      "Lovejoy"
## [16] "Ellicott"           "Pratt-Willert"      "Seneca Babcock"
## [19] "South Park"         "First Ward"         ""
## [22] "Hopkins-Tifft"      "Kaisertown"         "Seneca-Cazenovia"
## [25] "North Park"         "Riverside"          "West Hertel"
## [28] "Black Rock"         "Grant-Amherst"      "Upper West Side"
## [31] "UNKNOWN"            "University Heights" "Kensington-Bailey"
## [34] "Fillmore-Leroy"     "Parkside"           "Kenfield"
## [37] "Central Park"
```

Create a table that shows median property value by neighborhood

```
tapply(myData2$TOTAL.VALUE,c(as.factor(myData2$NEIGHBORHOOD)),median)
```

```
##                          Allentown         Black Rock  Broadway Fillmore
##             233500             264500              54000              11000
##            Central       Central Park     Delavan Grider           Ellicott
##             400000             211000              36000              39000
##    Elmwood Bidwell     Elmwood Bryant     Fillmore-Leroy         First Ward
##             266000             238000              40000              39000
##         Fruit Belt    Genesee-Moselle      Grant-Amherst        Hamlin Park
##              29000              20000              73000              58000
##      Hopkins-Tifft         Kaisertown           Kenfield  Kensington-Bailey
##              75000              74000              52000              49000
##            Lovejoy    Lower West Side        Masten Park           MLK Park
##              55000             156000              32000              25000
##         North Park           Parkside      Pratt-Willert          Riverside
##             191000             260000              46000              73000
##      Schiller Park   Seneca-Cazenovia     Seneca Babcock         South Park
##              41000              81000              36500             126000
## University Heights            UNKNOWN    Upper West Side        West Hertel
##              89000              76000              84000              86000
##          West Side
##             121500
```