Explore the Toyota Corolla Data Set Using R and R Studio

Didem Bulut Aykurt

MIS510-1 – Data Mining and Visualization

Colorado State University-Global Campus

Dr. Emmanuel Tsukerman

Now 20, 2022

Uploading and importing the Toyota Corolla CSV file into RStudio

When do I start to learn new knowledge that comes up with why I need that tool? Where will I use it? What kind of problem would I solve? Why do most companies work with them for the R program? First, I would like to give some minor information about all my questions. The best reason to use R is that over 12K packages and libraries are available and handle large amounts of unstructured data, all for free or less than other programs. Also, safe and security are the reasons all companies need them. Let's look at my work; R has an analysis tool to create graphs and charts with big and heavy data. It has an excellent console to manage code, data, and results and display graphs and charts. I download the R program, then the RStudio software project. Open the RStudio; go File, New File, R Script, then new page pop up.

I worked with the ToyotaCorolla.xlsx dataset. I convert the xlsx file to a CSV file with excel. Load the file into R Studio with <- read.csv("file path"), header=TRUE (if the dataset has a True header)).

Hence, let's dive deep into the Toyota Corolla dataset with the R programing language with RStudio. The ToyotaCorolla dataset contains car detail information that helps to compare car quality-related prices. This dataset is available from CSUGlobal. The function of "dim()" shows dataset dimension as the Toyota corolla dataset includes 1436 observations with 39 variables such as Id, Model, Price, Age_08_04, Mfg_Month, Mgf_Year, KM, Fuel_Type, HP, Met_Color, Color, Automatic, CC, Doors, Cylinders, Gears, Quarterly_Tax, Weight, Mfr_Guarantee, BOVAG_Guarantee, Guarantee_Period, ABS, Airbag_1, Airbag_2, Airco, Automatic_airco, Boardcomputer, CD_Player, Central_Lock, Power_Windows, Power_Steering, Radio, Mistlamps, Sport_Model,

Backseat_divider, Metallic_Rim, Radio_cassette, Parking_Assistant, Tow_Bar. The syntax of "view()" shows the whole table nicely on the other page. The "head()" shows all variables with six rows. The "mean()" and "summary()" display statistical results. When looking at the Price statistical result mean is the average number on the dataset of $10,731, the median is the middle number on the data point of $9,900, and the max is $32,500. Using the sample data is an easy and fast way to compare results between samples. R has a syntax is "sample <- sample(row. names(dataset name), number of observation) to create a random sample. Each variable has its category with "class(variable name)" and levels with "unique(variable names), a new name with "colnames(dataset name) <-c("put new column name"). Use the "model.matrix()" to create dummy table.

Concerns:

I am facing an issue with implementing complex code in R, like textbooks show levels() syntax that doesn't work with the Toyota corolla dataset; why the levels() function didn't work with the dataset?

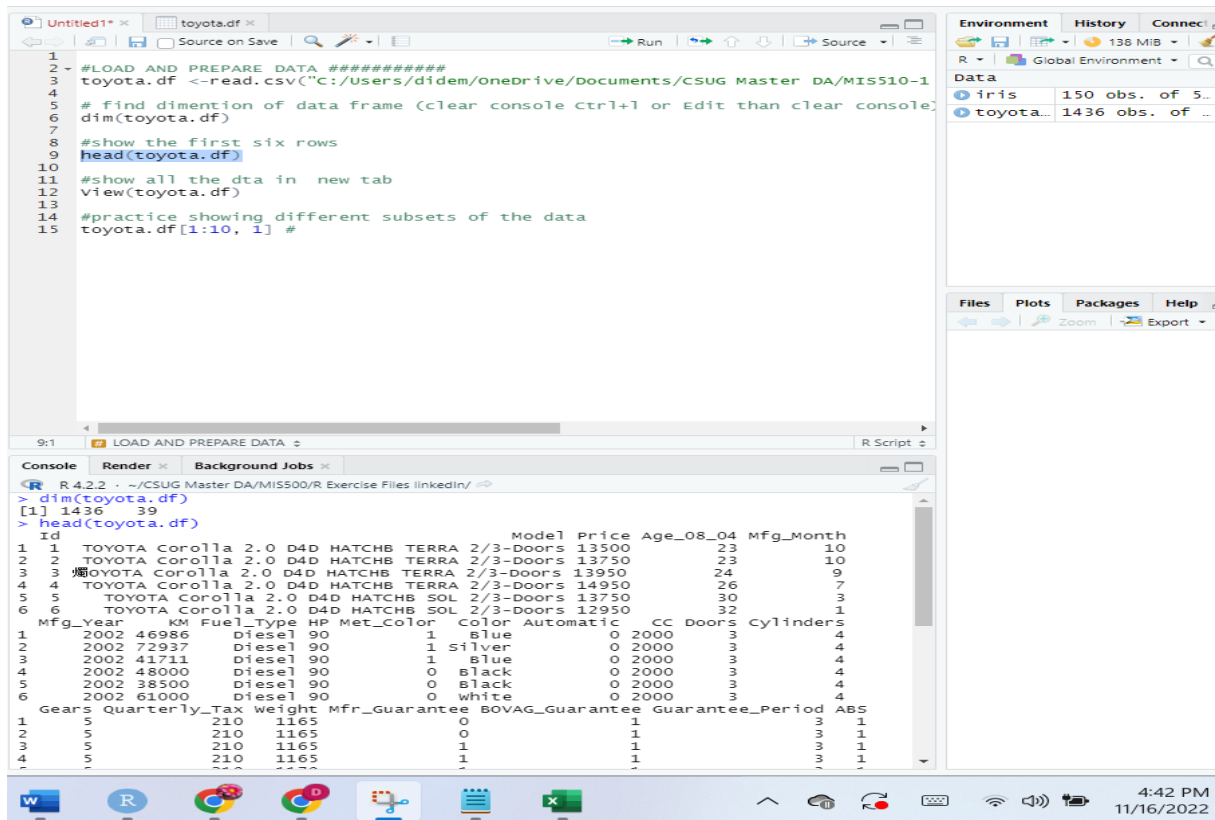**Figure 1:** *R code to import toyotacorolla.csv file into RStudio and explore the dataset.*

Figure 2: *R code as a "view ()" syntax result table.*

Figur3: *The R code display different subsets of the Toyota corolla dataset.*

◁▷ | 🗗 | 🔚 ☐ Source on Save | 🔍 🖉 ▾ | 🗏   → Run | ↱ ⇧ ⇩ | → Source ▾ ≡

```
1
2 ▾ #LOAD AND PREPARE DATA ###########
3   toyota.df <-read.csv("C:/Users/didem/OneDrive/Documents/CSUG Master DA/MIS510-1 Data Mining_4 term/Moduel
4
5   # find dimention of data frame (clear console Ctrl+l or Edit than clear console)
6   dim(toyota.df)
7
8   #show the first six rows
9   head(toyota.df)
10
11  #show all the dta in  new tab
12  View(toyota.df)
13
14  #practice showing different subsets of the data
15
16  #display the first column with first 10 row
17  toyota.df[1:10, 1]
18
19  #display the first 10 rows of each columns
20  toyota.df[1:10, ]
21
22  |
23
```

22:1 | ⊞ LOAD AND PREPARE DATA ‡   R Script

Console **Render** × **Background Jobs** ×   ▬ ☐

R R 4.2.2 · ~/CSUG Master DA/MIS500/R Exercise Files linkedIn/ ⇗

```
> toyota.df[1:10, 1]
 [1]  1  2  3  4  5  6  7  8  9 10
> toyota.df[1:10, ]
   Id                                    Model Price Age_08_04 Mfg_Month
1   1  TOYOTA Corolla 2.0 D4D HATCHB TERRA 2/3-Doors 13500      23       10
2   2  TOYOTA Corolla 2.0 D4D HATCHB TERRA 2/3-Doors 13750      23       10
3   3 燭OYOTA Corolla 2.0 D4D HATCHB TERRA 2/3-Doors 13950      24        9
4   4  TOYOTA Corolla 2.0 D4D HATCHB TERRA 2/3-Doors 14950      26        7
5   5    TOYOTA Corolla 2.0 D4D HATCHB SOL 2/3-Doors 13750      30        3
6   6    TOYOTA Corolla 2.0 D4D HATCHB SOL 2/3-Doors 12950      32        1
7   7 燭OYOTA Corolla 2.0 D4D 90 3DR TERRA 2/3-Doors 16900      27        6
8   8  TOYOTA Corolla 2.0 D4D 90 3DR TERRA 2/3-Doors 18600      30        3
9   9   燭OYOTA Corolla 1800 T SPORT VVT I 2/3-Doors 21500      27        6
10 10   燭OYOTA Corolla 1.9 D HATCHB TERRA 2/3-Doors 12950      23       10
   Mfg_Year     KM Fuel_Type  HP Met_Color  Color Automatic   CC Doors Cylinders
1      2002 46986    Diesel   90         1   Blue         0 2000     3         4
2      2002 72937    Diesel   90         1 Silver         0 2000     3         4
3      2002 41711    Diesel   90         1   Blue         0 2000     3         4
4      2002 48000    Diesel   90         0  Black         0 2000     3         4
5      2002 38500    Diesel   90         0  Black         0 2000     3         4
6      2002 61000    Diesel   90         0  White         0 2000     3         4
7      2002 94612    Diesel   90         1   Grey         0 2000     3         4
```

W R 🍩 🌀D ✂ ▤ X   ∧ ⬡ ↻ ⌨ 🛜 🔊 🔋   4:49 PM
                                                                   11/16/2022

```
1
2  #show the fifth row of first 10 columns
3  toyota.df[5, 1:10]
4
5  #show the fifth row of some columns
6  toyota.df[5, c(1:2, 4, 8:10)]
7
```

:30   # LOAD AND PREPARE DATA ⬍   R Script ⬍

sole   Render ×   Background Jobs ×   ▬ ☐

R 4.2.2 · ~/CSUG Master DA/MIS500/R Exercise Files linkedIn/

```
toyota.df[5, 1:10]
td                                  Model Price Age_08_04 Mfg_Month Mfg_Year    KM Fuel_Type HP
5 TOYOTA Corolla 2.0 D4D HATCHB SOL 2/3-Doors 13750        30         3    2002 38500    Diesel 90
Met_Color
       0
toyota.df[5, c(1:2, 4, 8:10)]
td                                  Model Age_08_04 Fuel_Type HP Met_Color
5 TOYOTA Corolla 2.0 D4D HATCHB SOL 2/3-Doors        30    Diesel 90         0
```

```
27
28  #show the whole first colum
29  toyota.df[ , 1]
30
```

9:1   # LOAD AND PREPARE DATA ⬍   R Script ⬍

sole   Render ×   Background Jobs ×   ▢▢

R 4.2.2 · ~/CSUG Master DA/MIS500/R Exercise Files linkedln/

```
toyota.df[ , 1]
  [1]    1    2    3    4    5    6    7    8    9   10   11   12   13   14   15   16   17   18   19   20
 [21]   21   22   23   24   25   26   27   28   29   30   31   32   33   34   35   36   37   38   39   40
 [41]   41   42   43   44   45   46   47   48   49   50   51   52   53   54   55   56   57   58   59   60
 [61]   61   62   63   64   65   66   67   68   69   70   71   72   73   74   75   76   77   78   79   80
 [81]   81   82   83   84   85   86   87   88   89   90   91   92   93   94   95   96   97   98   99  100
[101]  101  102  103  104  105  106  107  108  109  111  112  113  114  115  116  117  118  119  120  121
[121]  122  123  124  125  126  127  128  129  130  131  132  133  134  135  136  137  138  139  140  141
[141]  142  143  144  145  146  147  148  149  150  151  152  153  154  155  156  157  158  159  160  161
[161]  162  163  164  165  166  167  168  169  170  171  172  173  174  175  176  177  178  179  180  181
[181]  182  183  184  185  186  187  188  189  190  191  192  193  194  195  196  197  198  199  200  201
[201]  202  203  204  205  206  207  208  209  210  211  212  213  214  215  216  217  218  219  220  221
[221]  222  223  224  225  226  227  228  229  230  231  232  233  234  235  236  237  238  239  240  241
[241]  242  243  244  245  246  247  248  249  250  251  252  253  254  255  256  257  258  259  260  261
[261]  262  263  264  265  266  267  268  269  270  271  272  273  274  275  276  277  278  279  280  281
[281]  282  283  284  285  286  287  288  289  290  291  292  293  294  295  296  297  298  299  300  301
[301]  302  303  304  305  306  307  308  309  310  311  312  313  314  315  316  317  318  319  320  321
[321]  322  323  324  325  326  327  328  329  330  331  332  333  334  335  336  337  339  340  341  342
[341]  343  344  345  346  347  348  349  350  351  352  353  354  355  356  357  358  359  360  361  362
[361]  363  364  365  366  367  368  369  370  371  372  373  374  375  376  377  378  379  380  381  382
[381]  383  384  385  386  387  388  389  390  391  392  393  394  395  396  397  398  399  400  401  402
[401]  403  404  405  406  407  408  409  410  411  412  413  414  415  416  417  418  419  420  421  422
```

4:57 PM
11/16/2022

```
1  #different way to show the whole first column
2  toyota.df$Id
3
4  #show the first 10 rows of the first column
5  toyota.df$Id[1:10]
6
7
```

2:1   # LOAD AND PREPARE DATA ⬍   R Script ⬍

sole   Render ×   Background Jobs ×   ▢▢

R 4.2.2 · ~/CSUG Master DA/MIS500/R Exercise Files linkedln/

```
[721]  724  725  726  727  728  729  730  731  732  733  734  735  736  737  738  739  740  741
[739]  742  743  744  745  746  747  748  749  750  751  752  753  754  755  756  757  758  759
[757]  760  761  762  763  764  765  766  767  768  769  770  771  772  773  774  775  776  777
[775]  778  779  780  781  782  783  784  785  786  787  788  789  790  791  792  793  794  795
[793]  796  797  798  799  800  801  802  803  804  805  806  807  808  809  810  811  812  813
[811]  814  815  816  817  818  819  820  821  822  823  824  825  826  827  828  829  830  831
[829]  832  833  834  835  836  837  838  839  840  841  842  843  844  845  846  847  848  849
[847]  850  851  852  853  854  855  856  857  858  859  860  861  862  863  864  865  866  867
[865]  868  869  870  871  872  873  874  875  876  877  878  879  880  881  882  883  884  885
[883]  886  887  888  889  890  891  892  893  894  895  896  897  898  899  900  901  902  903
[901]  904  905  906  907  908  909  910  911  912  913  914  915  916  917  918  919  920  921
[919]  922  923  924  925  926  927  928  929  930  931  932  933  934  935  936  937  938  939
[937]  940  941  942  943  944  945  946  947  948  949  950  951  952  953  954  955  956  957
[955]  958  959  960  961  962  963  964  965  966  967  968  969  970  972  973  974  975  976
[973]  977  978  979  980  981  982  983  984  985  986  987  988  989  990  991  992  993  994
[991]  995  996  997  998  999 1000 1001 1002 1003 1004
 reached getOption("max.print") -- omitted 436 entries ]

#show the first 10 rows of the first column
toyota.df$Id[1:10]
[1]  1  2  3  4  5  6  7  8  9 10
```

6:04 PM
11/16/2022

Figure 4: *R code result of statistical information with Toyota corolla dataset.*

Figure 5: *R code creates a sample dataset from the database.*

Figure 6: *R code result of types of variables.*

```
54  #print a list of variables to the screen
55  names(toyota.df)
```

55:1   # LOAD AND PREPARE DATA ↕   R Script ↕

Console   Render ×   Background Jobs ×

R 4.2.2 · ~/CSUG Master DA/MIS500/R Exercise Files linkedIn/

```
> names(toyota.df)
 [1] "Id"                "Model"             "Price"             "Age_08_04"
 [5] "Mfg_Month"         "Mfg_Year"          "KM"                "Fuel_Type"
 [9] "HP"                "Met_Color"         "Color"             "Automatic"
[13] "CC"                "Doors"             "Cylinders"         "Gears"
[17] "Quarterly_Tax"     "Weight"            "Mfr_Guarantee"     "BOVAG_Guarantee"
[21] "Guarantee_Period"  "ABS"               "Airbag_1"          "Airbag_2"
[25] "Airco"             "Automatic_airco"   "Boardcomputer"     "CD_Player"
[29] "Central_Lock"      "Powered_Windows"   "Power_Steering"    "Radio"
[33] "Mistlamps"         "Sport_Model"       "Backseat_Divider"  "Metallic_Rim"
[37] "Radio_cassette"    "Parking_Assistant" "Tow_Bar"
>
```

6:23 PM
11/16/2022

```
57  #print the list in a useful column format
58  t(t(names(toyota.df)))
```

8:1   # LOAD AND PREPARE DATA ↕   R Script ↕

sole   Render ×   Background Jobs ×

R 4.2.2 · ~/CSUG Master DA/MIS500/R Exercise Files linkedIn/

```
:(t(names(toyota.df)))
      [,1]
1,] "Id"
2,] "Model"
3,] "Price"
4,] "Age_08_04"
5,] "Mfg_Month"
5,] "Mfg_Year"
7,] "KM"
8,] "Fuel_Type"
9,] "HP"
),] "Met_Color"
1,] "Color"
2,] "Automatic"
3,] "CC"
4,] "Doors"
5,] "Cylinders"
5,] "Gears"
7,] "Quarterly_Tax"
8,] "Weight"
9,] "Mfr_Guarantee"
),] "BOVAG_Guarantee"
```

6:27 PM
11/16/2022

```
60  # change first column's name
61  colnames(toyota.df)[1] <-c("Row Number")
62  names(toyota.df)[1]
```

61:1   # LOAD AND PREPARE DATA ÷   R Script ÷

nsole   Render ×   Background Jobs ×

R 4.2.2 · ~/CSUG Master DA/MIS500/R Exercise Files linkedIn/

```
3,]  "Airbag_1"
4,]  "Airbag_2"
5,]  "Airco"
6,]  "Automatic_airco"
7,]  "Boardcomputer"
8,]  "CD_Player"
9,]  "Central_Lock"
0,]  "Powered_Windows"
1,]  "Power_Steering"
2,]  "Radio"
3,]  "Mistlamps"
4,]  "Sport_Model"
5,]  "Backseat_Divider"
6,]  "Metallic_Rim"
7,]  "Radio_cassette"
8,]  "Parking_Assistant"
9,]  "Tow_Bar"
colnames(toyota.df)[1] <-c("Row Number")
colnames(toyota.df)[1] <-c("Row Number")
names(toyota.df)[1]
] "Row Number"
```

6:32 PM
11/16/2022

```
64  #Price is a integer variable
65  class(toyota.df$Price)
66
67  #show the different way variable category
68  class(toyota.df[ ,8])
69
70  #it can take one of levels
71
72  unique(toyota.df[,8])
73
74  #Model is a character variable
75  class(toyota.df$Model)
76
77
```

65:23   # LOAD AND PREPARE DATA ÷   R Script ÷

Console   Render ×   Background Jobs ×

R 4.2.2 · ~/CSUG Master DA/MIS500/R Exercise Files linkedIn/

```
Mean   :0.1455   Mean   :0.002786   Mean   :0.2779
3rd Qu.:0.0000   3rd Qu.:0.000000   3rd Qu.:1.0000
Max.   :1.0000   Max.   :1.000000   Max.   :1.0000
> class(toyota.df$Model)
[1] "character"
> #Price is a integer variable
> class(toyota.df$Price)
[1] "integer"
>
> #show the different way variable category
> class(toyota.df[ ,8])
[1] "character"
>
> #it can take one of levels
>
> unique(toyota.df[,8])
[1] "Diesel" "Petrol" "CNG"
>
> #Model is a character variable
> class(toyota.df$Model)
[1] "character"
> |
```

Files   Plots   Packages   Help

Install   Update   🔍

Na...	Description	...
	to MIME Types	
ml...	Machine Learning Benchmark Problems	2.1· 3
mlr	Machine Learning in R	2.19
mlr3	Machine Learning in R - Next Generation	0.13
mlr...	Performance Measures for 'mlr3'	0.4.
mlr...	Helper Functions for 'mlr3'	0.10
Mo...	Rapid Calculation of Model Metrics	1.2.
mo...	Modelling Functions that Work with the Pipe	0.1.
nlme	Linear and Nonlinear Mixed Effects Models	3.1· 160
nnet	Feed-Forward Neural	7.3· 18

1:53 PM
11/18/2022

Figure 7: *Create a dummy table with the Toyota corolla dataset in R code.*