Subsets of Data Set

Code

Vanessa Gonzalez

2018-06-25

CS Graduated Students Data Set

Hide

```
library("caret")
dfDataSet <- as.data.frame(dfDataSet)
summary(dfDataSet)
```

```
 Year of OriginalMajorDate       GraduationStatus  YearsFromOMD          CsGrad     4
YG        5YG        6YG         1_CSCI101
 Min.   :2008              CurrentStudent: 32     Min.   :3.830   NG         :140   No
:252   No :169   No :150   Min.   :0.300
 1st Qu.:2009              Graduated     :396     1st Qu.:4.830   OtherMajor: 57   Ye
s:284   Yes:367   Yes:386   1st Qu.:3.000
 Median :2011              InactiveReg   :108     Median :6.840   Yes        :339
Median :4.000
 Mean   :2011                                    Mean   :6.706
Mean   :3.419
 3rd Qu.:2013                                    3rd Qu.:8.840
3rd Qu.:4.000
 Max.   :2014                                    Max.   :9.840
Max.   :4.000

NA's   :76
   1_MATH111      2_CSCI261       2_MATH112       2_MATH201       3_CSCI262       3_MA
TH213        4_CSCI341       4_CSCI358
 Min.   :0.30   Min.   :0.300   Min.   :0.300   Min.   :0.300   Min.   :0.30   Min.
:0.300   Min.   :0.300   Min.   :0.300
 1st Qu.:3.00   1st Qu.:3.000   1st Qu.:2.000   1st Qu.:2.000   1st Qu.:3.00   1st Qu
.:2.000   1st Qu.:2.000   1st Qu.:2.000
 Median :3.00   Median :4.000   Median :3.000   Median :3.000   Median :4.00   Median
:3.000   Median :3.000   Median :3.000
 Mean   :2.91   Mean   :3.405   Mean   :2.875   Mean   :2.701   Mean   :3.24   Mean
:2.888   Mean   :2.854   Mean   :2.961
 3rd Qu.:3.00   3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:3.300   3rd Qu.:4.00   3rd Qu
.:4.000   3rd Qu.:4.000   3rd Qu.:4.000
 Max.   :4.00   Max.   :4.000   Max.   :4.000   Max.   :4.000   Max.   :4.00   Max.
:4.000   Max.   :4.000   Max.   :4.000
```

```
 NA's   :12     NA's   :37      NA's   :26      NA's   :108     NA's   :81     NA's
:49      NA's   :108     NA's   :108
   4_MATH225       5_CSCI306       5_CSCI403       5_MATH332       6_CSCI406       7_
CSCI370       8_CSCI400       9_CSCI442
 Min.   :0.300   Min.   :0.300   Min.   :0.300   Min.   :0.300   Min.   :0.300   Min.
:2.300   Min.   :0.300   Min.   :0.300
 1st Qu.:2.000   1st Qu.:3.000   1st Qu.:3.000   1st Qu.:2.000   1st Qu.:2.000   1st
Qu.:4.000   1st Qu.:2.925   1st Qu.:3.000
 Median :3.000   Median :3.700   Median :4.000   Median :3.000   Median :3.000   Medi
an :4.000   Median :3.300   Median :3.300
 Mean   :2.763   Mean   :3.433   Mean   :3.565   Mean   :2.671   Mean   :2.795   Mean
:3.895   Mean   :3.175   Mean   :3.133
 3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:3.300   3rd Qu.:3.300   3rd
Qu.:4.000   3rd Qu.:4.000   3rd Qu.:4.000
 Max.   :4.000   Max.   :4.000   Max.   :4.000   Max.   :4.000   Max.   :4.000   Max.
:4.000   Max.   :4.000   Max.   :4.000
 NA's   :58      NA's   :132     NA's   :250     NA's   :121     NA's   :154     NA's
:179     NA's   :164     NA's   :165
```

Create a subset of data consisting of students with a “GraduationStatus” of “Graduated”

Hide

```
GraduatedData<-subset(dfDataSet, GraduationStatus == 'Graduated')
head(GraduatedData)
```

	Year of OriginalMajorDate <int>	GraduationStatus <fctr>	YearsFromO... <dbl>	CsGrad <fctr>	4... <fctr>	5... <fctr>	6... <fctr
1	2014	Graduated	4.00	Yes	Yes	Yes	Yes
2	2008	Graduated	9.84	OtherMajor	No	Yes	Yes
5	2008	Graduated	9.84	OtherMajor	Yes	Yes	Yes
6	2008	Graduated	9.84	Yes	Yes	Yes	Yes
7	2008	Graduated	9.84	Yes	Yes	Yes	Yes
8	2008	Graduated	9.84	OtherMajor	Yes	Yes	Yes

6 rows | 1-9 of 24 columns

Look at the data subset

Hide

```
summary(GraduatedData)
```

```
 Year of OriginalMajorDate       GraduationStatus  YearsFromOMD          CsGrad     4
YG       5YG       6YG        1_CSCI101
 Min.   :2008              CurrentStudent:  0     Min.   :3.830   NG        :  0   No
:112   No : 29   No : 10   Min.   :0.300
 1st Qu.:2009              Graduated     :396     1st Qu.:4.830   OtherMajor: 57   Ye
s:284   Yes:367   Yes:386   1st Qu.:3.000
 Median :2011              InactiveReg   :  0     Median :6.840   Yes       :339
Median :4.000
 Mean   :2011                                    Mean   :6.885
Mean   :3.644
 3rd Qu.:2013                                    3rd Qu.:8.840
3rd Qu.:4.000
 Max.   :2014                                    Max.   :9.840
Max.   :4.000

NA's   :43
   1_MATH111       2_CSCI261       2_MATH112       2_MATH201       3_CSCI262       3_
MATH213       4_CSCI341       4_CSCI358
 Min.   :1.000   Min.   :0.300   Min.   :0.300   Min.   :0.700   Min.   :0.300   Min.
:1.000   Min.   :0.300   Min.   :0.700
 1st Qu.:3.000   1st Qu.:3.000   1st Qu.:3.000   1st Qu.:2.000   1st Qu.:3.000   1st
Qu.:2.000   1st Qu.:2.000   1st Qu.:2.300
 Median :3.000   Median :4.000   Median :3.000   Median :3.000   Median :4.000   Medi
an :3.000   Median :3.000   Median :3.000
 Mean   :3.033   Mean   :3.583   Mean   :3.037   Mean   :2.825   Mean   :3.475   Mean
:3.055   Mean   :3.097   Mean   :3.092
 3rd Qu.:3.000   3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:3.700   3rd Qu.:4.000   3rd
Qu.:4.000   3rd Qu.:4.000   3rd Qu.:4.000
 Max.   :4.000   Max.   :4.000   Max.   :4.000   Max.   :4.000   Max.   :4.000   Max.
:4.000   Max.   :4.000   Max.   :4.000
 NA's   :7       NA's   :11      NA's   :6       NA's   :32      NA's   :33      NA's
:6       NA's   :45      NA's   :39
   4_MATH225       5_CSCI306       5_CSCI403       5_MATH332       6_CSCI406       7_
CSCI370       8_CSCI400       9_CSCI442
 Min.   :1.000   Min.   :0.300   Min.   :0.300   Min.   :0.700   Min.   :0.300   Min.
:2.700   Min.   :0.700   Min.   :0.700
 1st Qu.:2.000   1st Qu.:3.000   1st Qu.:4.000   1st Qu.:2.000   1st Qu.:2.000   1st
Qu.:4.000   1st Qu.:3.000   1st Qu.:3.000
 Median :3.000   Median :3.700   Median :4.000   Median :3.000   Median :3.000   Medi
an :4.000   Median :3.700   Median :3.300
 Mean   :2.929   Mean   :3.501   Mean   :3.669   Mean   :2.817   Mean   :2.888   Mean
:3.909   Mean   :3.288   Mean   :3.231
 3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:3.700   3rd Qu.:4.000   3rd
Qu.:4.000   3rd Qu.:4.000   3rd Qu.:4.000
 Max.   :4.000   Max.   :4.000   Max.   :4.000   Max.   :4.000   Max.   :4.000   Max.
:4.000   Max.   :4.000   Max.   :4.000
 NA's   :4       NA's   :48      NA's   :148     NA's   :47      NA's   :53      NA's
```

```
:58      NA's   :59      NA's   :57
```

Hide

```
str(GraduatedData)
```

```
'data.frame':   396 obs. of  24 variables:
 $ Year of OriginalMajorDate: int  2014 2008 2008 2008 2008 2008 2008 2008 2008 2008
...
 $ GraduationStatus         : Factor w/ 3 levels "CurrentStudent",..: 2 2 2 2 2 2 2 2
2 2 ...
 $ YearsFromOMD             : num  4 9.84 9.84 9.84 9.84 9.84 9.84 9.84 9.84 9.84 ...
 $ CsGrad                   : Factor w/ 3 levels "NG","OtherMajor",..: 3 2 2 3 3 2 2
3 3 3 ...
 $ 4YG                      : Factor w/ 2 levels "No","Yes": 2 1 2 2 2 2 2 2 2 2 ...
 $ 5YG                      : Factor w/ 2 levels "No","Yes": 2 2 2 2 2 2 2 2 2 2 ...
 $ 6YG                      : Factor w/ 2 levels "No","Yes": 2 2 2 2 2 2 2 2 2 2 ...
 $ 1_CSCI101                : num  4 NA NA 4 4 NA NA 4 4 3 ...
 $ 1_MATH111                : num  3 3 3 3 4 3 3 3 2 4 ...
 $ 2_CSCI261                : num  4 4 4 3 4 3 3 4 3 4 ...
 $ 2_MATH112                : num  2 2 3 3 4 4 2 3 3 3 ...
 $ 2_MATH201                : num  3 3 3 2 4 NA 1 3 2 2 ...
 $ 3_CSCI262                : num  3 NA NA 3 4 4 NA 4 4 3 ...
 $ 3_MATH213                : num  4 3 3 4 4 4 2 3 4 3 ...
 $ 4_CSCI341                : num  2 NA NA 3 4 NA 3 4 3 3 ...
 $ 4_CSCI358                : num  4 NA NA 2 4 NA NA 4 4 3 ...
 $ 4_MATH225                : num  4 3 4 3 4 4 1 4 4 3 ...
 $ 5_CSCI306                : num  3 NA NA 4 4 4 NA 4 4 4 ...
 $ 5_CSCI403                : num  4 NA NA 4 4 NA NA NA 4 4 ...
 $ 5_MATH332                : num  3 NA NA 3 4 NA NA 2 4 3 ...
 $ 6_CSCI406                : num  2 NA NA 2 4 NA NA 4 3 3 ...
 $ 7_CSCI370                : num  3.3 NA NA 4 4 NA NA 4 4 4 ...
 $ 8_CSCI400                : num  3.3 NA NA 3 4 NA NA 4 4 3 ...
 $ 9_CSCI442                : num  2.3 NA NA 3 4 NA NA 4 4 3 ...
```

Remove not needed columns from data set and leave factor Four-year Graduation Factor

Hide

```
DataSet4YG <- GraduatedData[(5:24)]
DataSet4YG <- DataSet4YG[-(2:3)]
head(DataSet4YG)
```

4...	1_CSCI1...	1_MATH...	2_CSCI2...	2_MATH...	2_MATH...	3_CSCI2...	3_MATH...	4_CSCI3.
<fctr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl

1	Yes	4	3	4	2	3	3	4	
2	No	*NA*	3	4	2	3	*NA*	3	*N*
5	Yes	*NA*	3	4	3	3	*NA*	3	*N*
6	Yes	4	3	3	3	2	3	4	
7	Yes	4	4	4	4	4	4	4	
8	Yes	*NA*	3	3	4	*NA*	4	4	*N*

6 rows | 1-10 of 18 columns

Transform data set into a data frame

Hide

```
dfDataSet4YG <-as.data.frame(DataSet4YG)
```

Find coorelation between variables using “spearman” method

Hide

```
res<- cor(dfDataSet4YG[-(1)], method = 'spearman', use = "complete.obs")
round(res,2)
```

```
          1_CSCI101 1_MATH111 2_CSCI261 2_MATH112 2_MATH201 3_CSCI262 3_MATH213 4_CSC
I341 4_CSCI358 4_MATH225 5_CSCI306 5_CSCI403
1_CSCI101      1.00      0.11      0.26      0.36      0.30      0.40      0.29
0.27      0.21      0.36      0.27      0.14
1_MATH111      0.11      1.00      0.22      0.42      0.19      0.04      0.34
0.19      0.12      0.26      0.18     -0.04
2_CSCI261      0.26      0.22      1.00      0.19      0.23      0.26      0.20
0.29      0.15      0.26      0.23      0.09
2_MATH112      0.36      0.42      0.19      1.00      0.45      0.29      0.47
0.35      0.36      0.42      0.34      0.11
2_MATH201      0.30      0.19      0.23      0.45      1.00      0.44      0.45
0.55      0.45      0.59      0.54      0.31
3_CSCI262      0.40      0.04      0.26      0.29      0.44      1.00      0.37
0.50      0.32      0.45      0.45      0.32
3_MATH213      0.29      0.34      0.20      0.47      0.45      0.37      1.00
0.47      0.42      0.57      0.33      0.22
4_CSCI341      0.27      0.19      0.29      0.35      0.55      0.50      0.47
1.00      0.36      0.58      0.44      0.25
4_CSCI358      0.21      0.12      0.15      0.36      0.45      0.32      0.42
0.36      1.00      0.44      0.36      0.16
4_MATH225      0.36      0.26      0.26      0.42      0.59      0.45      0.57
0.58      0.44      1.00      0.47      0.31
```

```
5_CSCI306       0.27       0.18       0.23       0.34       0.54       0.45       0.33
0.44       0.36       0.47       1.00       0.35
5_CSCI403       0.14      -0.04       0.09       0.11       0.31       0.32       0.22
0.25       0.16       0.31       0.35       1.00
5_MATH332       0.33       0.20       0.26       0.47       0.61       0.56       0.55
0.54       0.46       0.62       0.53       0.33
6_CSCI406       0.32       0.16       0.20       0.38       0.58       0.48       0.41
0.43       0.42       0.50       0.55       0.34
7_CSCI370       0.13      -0.02       0.01       0.12       0.15       0.26       0.13
0.19       0.00       0.16       0.15       0.09
8_CSCI400       0.35       0.09       0.22       0.34       0.59       0.60       0.41
0.53       0.41       0.46       0.61       0.46
9_CSCI442       0.36       0.15       0.22       0.37       0.57       0.51       0.42
0.44       0.30       0.50       0.51       0.35
           5_MATH332 6_CSCI406 7_CSCI370 8_CSCI400 9_CSCI442
1_CSCI101       0.33       0.32       0.13       0.35       0.36
1_MATH111       0.20       0.16      -0.02       0.09       0.15
2_CSCI261       0.26       0.20       0.01       0.22       0.22
2_MATH112       0.47       0.38       0.12       0.34       0.37
2_MATH201       0.61       0.58       0.15       0.59       0.57
3_CSCI262       0.56       0.48       0.26       0.60       0.51
3_MATH213       0.55       0.41       0.13       0.41       0.42
4_CSCI341       0.54       0.43       0.19       0.53       0.44
4_CSCI358       0.46       0.42       0.00       0.41       0.30
4_MATH225       0.62       0.50       0.16       0.46       0.50
5_CSCI306       0.53       0.55       0.15       0.61       0.51
5_CSCI403       0.33       0.34       0.09       0.46       0.35
5_MATH332       1.00       0.59       0.04       0.57       0.54
6_CSCI406       0.59       1.00       0.11       0.59       0.64
7_CSCI370       0.04       0.11       1.00       0.28       0.16
8_CSCI400       0.57       0.59       0.28       1.00       0.56
9_CSCI442       0.54       0.64       0.16       0.56       1.00
```

To substitute NA values with another value the KNN Imputation method is used

Hide

```
library("DMwR")
DataSet4YGImpute <- knnImputation(DataSet4YG)
head(DataSet4YGImpute)
```

	4... <fctr>	1_CSCI1... <dbl>	1_MATH... <dbl>	2_CSCI2... <dbl>	2_MATH... <dbl>	2_MATH... <dbl>	3_CSCI2... <dbl>	3_MATH... <dbl>	4_CSCI3. <dbl
1	Yes	4.000000	3	4	2	3.000000	3.000000	4	2.00000
2	No	3.625546	3	4	2	3.000000	3.603932	3	3.10005

5	Yes	3.776819	3	4	3	3.000000	3.665798	3	3.66465
6	Yes	4.000000	3	3	3	2.000000	3.000000	4	3.00000
7	Yes	4.000000	4	4	4	4.000000	4.000000	4	4.00000
8	Yes	4.000000	3	3	4	3.790084	4.000000	4	3.51737

6 rows | 1-10 of 18 columns

Hide

```
str(DataSet4YGImpute)
```

```
'data.frame':   396 obs. of  18 variables:
 $ 4YG      : Factor w/ 2 levels "No","Yes": 2 1 2 2 2 2 2 2 2 2 ...
 $ 1_CSCI101: num  4 3.63 3.78 4 4 ...
 $ 1_MATH111: num  3 3 3 3 4 3 3 3 2 4 ...
 $ 2_CSCI261: num  4 4 4 3 4 3 3 4 3 4 ...
 $ 2_MATH112: num  2 2 3 3 4 4 2 3 3 3 ...
 $ 2_MATH201: num  3 3 3 2 4 ...
 $ 3_CSCI262: num  3 3.6 3.67 3 4 ...
 $ 3_MATH213: num  4 3 3 4 4 4 2 3 4 3 ...
 $ 4_CSCI341: num  2 3.1 3.66 3 4 ...
 $ 4_CSCI358: num  4 3.1 3.72 2 4 ...
 $ 4_MATH225: num  4 3 4 3 4 4 1 4 4 3 ...
 $ 5_CSCI306: num  3 3.62 3.74 4 4 ...
 $ 5_CSCI403: num  4 3.9 4 4 4 ...
 $ 5_MATH332: num  3 2.56 2.78 3 4 ...
 $ 6_CSCI406: num  2 2.83 3.16 2 4 ...
 $ 7_CSCI370: num  3.3 4 4 4 4 ...
 $ 8_CSCI400: num  3.3 3.58 3.34 3 4 ...
 $ 9_CSCI442: num  2.3 2.9 3.23 3 4 ...
```

Hide

```
dfDataSet4YGImpute <-as.data.frame(DataSet4YGImpute)
```

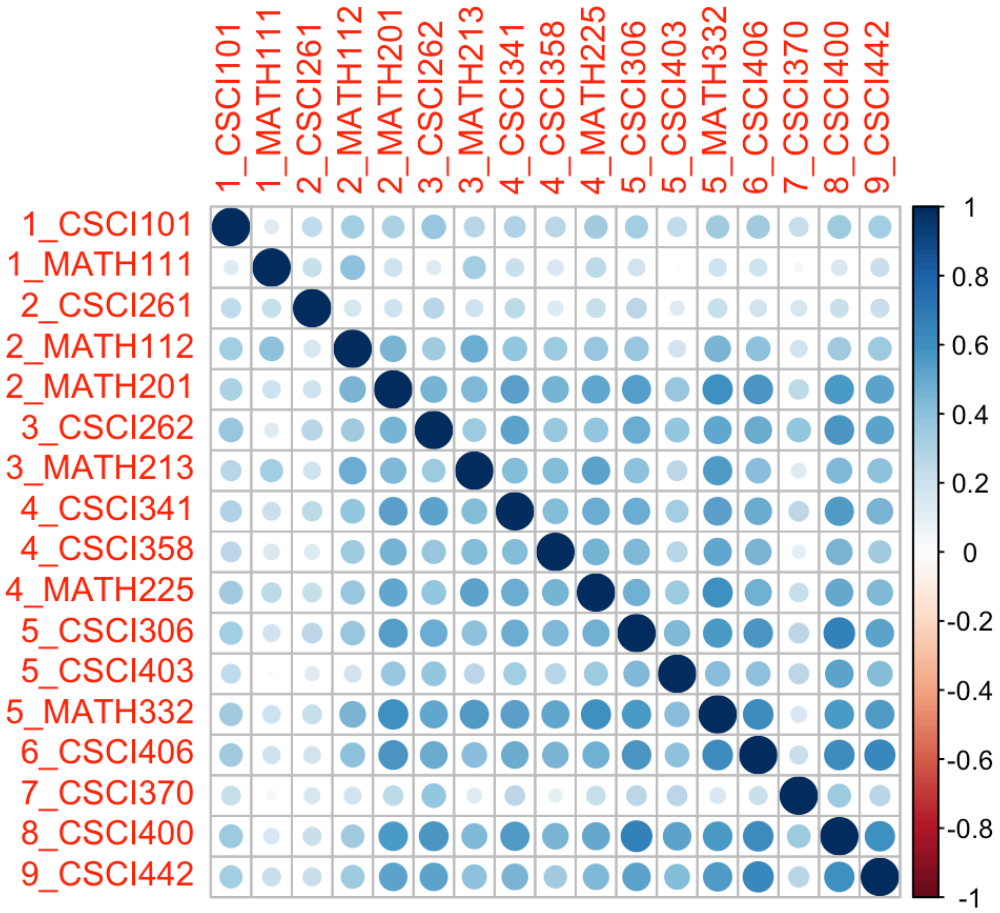
Create a correlation plot between variables

Hide

```
library(corrplot)
```

Hide

```
corrplot(cor(dfDataSet4YGImpute[-(1)], method = 'spearman', use = "complete.obs"))
```



To determine variables with a correlation higher than 0.5

Hide

```
highlyCorrelated <- findCorrelation(res, cutoff=0.5)
print(highlyCorrelated)
```

```
[1] 13 16  5 10 14 17
```