MATH2349 Semester 1, 2020

Code ▾

Assignment 1

Akshat Vijayvargia s3826627

DATA DESCRIPTION

This dataset is made to understand the factors and their influence on the marks of the United States’s middle school going childrens.Source of the dataset is- *[https://www.kaggle.com/spscientist/students-performance-in-exams (https://www.kaggle.com/spscientist/students-performance-in-exams)] This data has 1,000 observations and 8 variables.

READ/IMPORT DATA

We introduce the data to R, make a dataframe out of it and save that dataframe, the process as follows-

Hide

```
# Importing  the data
students_performance_in_exams_ <- read.csv("C:/Users/61422/OneDrive/Desktop/DW/stud
ents-performance-in-exams..csv")

# Saving the dataset as a dataframe
df1 <- data.frame(students_performance_in_exams_ )
print(df1)
```

gen… <fctr>	**race.ethnicity** <fctr>	**parental.level.of.education** <fctr>	**lunch** <fctr>	**test.prep** <fctr>
female	group B	bachelor's degree	standard	none
female	group C	some college	standard	complete
female	group B	master's degree	standard	none
male	group A	associate's degree	free/reduced	none
male	group C	some college	standard	none
female	group B	associate's degree	standard	none
female	group B	some college	standard	complete
male	group B	some college	free/reduced	none
male	group D	high school	free/reduced	complete
female	group B	high school	free/reduced	none

1-10 of 1,000 rows | 1-6 of 8 columns Previous **1** 2 3 4 5 6 ... 100 Next

Hide

```
# Checking the head of the dataset
head(df1, n = 5)
```

	gen…	race.ethnicity	parental.level.of.education	lunch	test.prepa
	<fctr>	<fctr>	<fctr>	<fctr>	<fctr>
1	female	group B	bachelor's degree	standard	none
2	female	group C	some college	standard	completed
3	female	group B	master's degree	standard	none
4	male	group A	associate's degree	free/reduced	none
5	male	group C	some college	standard	none

5 rows | 1-6 of 8 columns

Hide

```
# Checking the tail of the dataset
tail(df1, n = 5)
```

	gen…	race.ethnicity	parental.level.of.education	lunch	test.p
	<fctr>	<fctr>	<fctr>	<fctr>	<fctr>
996	female	group E	master's degree	standard	compl
997	male	group C	high school	free/reduced	none
998	female	group C	high school	free/reduced	compl
999	female	group D	some college	standard	compl
1000	female	group D	some college	free/reduced	none

5 rows | 1-6 of 8 columns

INSPECT AND UNDERSTAND

*All the codes with explaination are stated below-

Hide

```
# Checking the dimensions of dataframe
dim(df1)
```

```
[1] 1000    8
```

Hide

```
# Checking the column names
colnames(df1)
```

```
[1] "gender"                      "race.ethnicity"              "parental.level.of.
education"
[4] "lunch"                       "test.preparation.course"     "math.score"
[7] "reading.score"               "writing.score"
```

Hide

```
# Checking the datatype
str(df1)
```

```
'data.frame':   1000 obs. of  8 variables:
 $ gender                     : Factor w/ 2 levels "female","male": 1 1 1 2 2 1 1 2
2 1 ...
 $ race.ethnicity             : Factor w/ 5 levels "group A","group B",..: 2 3 2 1
3 2 2 2 4 2 ...
 $ parental.level.of.education: Factor w/ 6 levels "associate's degree",..: 2 5 4 1
5 1 5 5 3 3 ...
 $ lunch                      : Factor w/ 2 levels "free/reduced",..: 2 2 2 1 2 2 2
1 1 1 ...
 $ test.preparation.course    : Factor w/ 2 levels "completed","none": 2 1 2 2 2 2
1 2 1 2 ...
 $ math.score                 : int  72 69 90 47 76 71 88 40 64 38 ...
 $ reading.score              : int  72 90 95 57 78 83 95 43 64 60 ...
 $ writing.score              : int  74 88 93 44 75 78 92 39 67 50 ...
```

Hide

```
# Converting the inccorect variable datatypes accordingly
# gender
df1$gender <- factor(df1$gender)

# level of education
df1$parental.level.of.education <- factor(df1$parental.level.of.education)

# test.preparation.course
df1$test.preparation.course <- factor(df1$test.preparation.course)

# lunch
df1$lunch <- factor(df1$lunch)

# Summary of the dataset
summary(df1)
```

```
    gender    race.ethnicity     parental.level.of.education          lunch     tes
t.preparation.course
 female:518   group A: 89    associate's degree:222          free/reduced:355   com
pleted:358
 male  :482   group B:190    bachelor's degree :118          standard    :645   non
e     :642
              group C:319    high school       :196
              group D:262    master's degree   : 59
              group E:140    some college      :226
                             some high school  :179
   math.score     reading.score    writing.score
 Min.   :  0.00   Min.   : 17.00   Min.   : 10.00
 1st Qu.: 57.00   1st Qu.: 59.00   1st Qu.: 57.75
 Median : 66.00   Median : 70.00   Median : 69.00
 Mean   : 66.09   Mean   : 69.17   Mean   : 68.05
 3rd Qu.: 77.00   3rd Qu.: 79.00   3rd Qu.: 79.00
 Max.   :100.00   Max.   :100.00   Max.   :100.00
```

Hide

```
# Checking the levels of variables
levels(df1$gender)
```

```
[1] "female" "male"
```

Hide

```
levels(df1$parental.level.of.education)
```

```
[1] "associate's degree" "bachelor's degree"  "high school"        "master's degre
e"    "some college"
[6] "some high school"
```

Hide

```
levels(df1$test.preparation.course)
```

```
[1] "completed" "none"
```

Hide

```
levels(df1$lunch <- factor(df1$lunch))
```

```
[1] "free/reduced" "standard"
```

SUBSETTING I

To subset the dataframe to only 10 observations, we can type in the 1:10 in the square bracket.

Hide

```
#subseting df1
df2 <- df1[1:10,]
print(df2)
```

	gen...	race.ethnicity	parental.level.of.education	lunch	test.prep
	<fctr>	<fctr>	<fctr>	<fctr>	<fctr>
1	female	group B	bachelor's degree	standard	none
2	female	group C	some college	standard	complete
3	female	group B	master's degree	standard	none
4	male	group A	associate's degree	free/reduced	none
5	male	group C	some college	standard	none
6	female	group B	associate's degree	standard	none
7	female	group B	some college	standard	complete
8	male	group B	some college	free/reduced	none
9	male	group D	high school	free/reduced	complete
10	female	group B	high school	free/reduced	none

1-10 of 10 rows | 1-6 of 8 columns

Hide

```


#converting the dataframe into matrix
matrix1 <-as.matrix(df2)

#checking the class of matrix
class(matrix1)
```

```
[1] "matrix"
```

Hide

```
# structure of matrix
str(matrix1)
```

```
 chr [1:10, 1:8] "female" "female" "female" "male" "male" "female" "female" "male"
"male" "female" ...
 - attr(*, "dimnames")=List of 2
  ..$ : chr [1:10] "1" "2" "3" "4" ...
  ..$ : chr [1:8] "gender" "race.ethnicity" "parental.level.of.education" "lunch"
...
```

Hide

```
print(matrix1)
```

```
   gender   race.ethnicity parental.level.of.education lunch          test.preparat
ion.course math.score
1  "female" "group B"      "bachelor's degree"         "standard"     "none"
"72"
2  "female" "group C"      "some college"              "standard"     "completed"
"69"
3  "female" "group B"      "master's degree"           "standard"     "none"
"90"
4  "male"   "group A"      "associate's degree"        "free/reduced" "none"
"47"
5  "male"   "group C"      "some college"              "standard"     "none"
"76"
6  "female" "group B"      "associate's degree"        "standard"     "none"
"71"
7  "female" "group B"      "some college"              "standard"     "completed"
"88"
8  "male"   "group B"      "some college"              "free/reduced" "none"
"40"
9  "male"   "group D"      "high school"               "free/reduced" "completed"
"64"
10 "female" "group B"      "high school"               "free/reduced" "none"
"38"
   reading.score writing.score
1  "72"          "74"
2  "90"          "88"
3  "95"          "93"
4  "57"          "44"
5  "78"          "75"
6  "83"          "78"
7  "95"          "92"
8  "43"          "39"
9  "64"          "67"
10 "60"          "50"
```

SUBSETTING II

To subset the dataframe for only specific variable, then selecting the fisrt and last variable through c(1,8).

Hide

```
#subseting the first and last variable of df2
df3 <- df2[ ,c(1,8)]
print(df3)
```

	gender <fctr>	writing.score <int>
1	female	74
2	female	88
3	female	93
4	male	44
5	male	75
6	female	78
7	female	92
8	male	39
9	male	67
10	female	50

1-10 of 10 rows

Hide

```
# Saving it as a .RData
saveRDS(df3,"studentsandscores.rds")
```

CREATE A NEW DATA FRAME

Now we create a new dataframe using data.frame() function.

Hide

```
#creating a new data frame
NAMES= c('Anirudh','Akshat','Riya','Rainy','Sunny','Aditya','Anil','Angelo','Ale
x','Danny')
AGE= c(20L,21L,22L,23L,24L,25L,26L,27L,28L,29L)
RANK= c('4th','3rd','2nd','5th','6th','7th','1st','8th','9th','1oth')
df4 <-data.frame(NAMES,AGE,RANK)
print(df4)
```

NAMES <fctr>	AGE <int>	RANK <fctr>
Anirudh	20	4th
Akshat	21	3rd
Riya	22	2nd
Rainy	23	5th

NAMES	AGE	RANK
<fctr>	<int>	<fctr>
Sunny	24	6th
Aditya	25	7th
Anil	26	1st
Angelo	27	8th
Alex	28	9th
Danny	29	1oth

1-10 of 10 rows

Hide

```
#checking the head and tails of the dataframe
head(df4)
```

	NAMES	AGE	RANK
	<fctr>	<int>	<fctr>
1	Anirudh	20	4th
2	Akshat	21	3rd
3	Riya	22	2nd
4	Rainy	23	5th
5	Sunny	24	6th
6	Aditya	25	7th

6 rows

Hide

```
tail(df4)
```

	NAMES	AGE	RANK
	<fctr>	<int>	<fctr>
5	Sunny	24	6th
6	Aditya	25	7th
7	Anil	26	1st
8	Angelo	27	8th
9	Alex	28	9th

	NAMES <fctr>	AGE <int>	RANK <fctr>
10	Danny	29	1oth

6 rows

Hide

```
#checking the structure of variable and levels of ordinal variable
str(NAMES,)
```

```
 chr [1:10] "Anirudh" "Akshat" "Riya" "Rainy" "Sunny" "Aditya" "Anil" "Angelo" "Ale
x" "Danny"
```

Hide

```
# adding a new column in the above dataframe
TIMETAKEN= c(43.45,43.29,43.18,43.48,43.50,43.64,43.03,43.65,43.74,43.81)
df5 <-data.frame(df4,TIMETAKEN)
print(df5)
```

NAMES <fctr>	AGE <int>	RANK <fctr>	TIMETAKEN <dbl>
Anirudh	20	4th	43.45
Akshat	21	3rd	43.29
Riya	22	2nd	43.18
Rainy	23	5th	43.48
Sunny	24	6th	43.50
Aditya	25	7th	43.64
Anil	26	1st	43.03
Angelo	27	8th	43.65
Alex	28	9th	43.74
Danny	29	1oth	43.81

1-10 of 10 rows

Hide

```
#checking all the attributes of dataframe
attributes(df5)
```

```
$names
[1] "NAMES"     "AGE"       "RANK"      "TIMETAKEN"

$class
[1] "data.frame"

$row.names
 [1]  1  2  3  4  5  6  7  8  9 10
```

Hide

```
#checking the dimensions of dataframe
dim(df5)
```

```
[1] 10  4
```