CS 5135/6035 Learning Probabilistic Models

Lecture 2a: Dataframes in Julia

Gowtham Atluri

August 28, 2018

Dataframe

- What is a dataframe?
- How are they different from data matrices?
- What can we do with them?
- How do we use them to explore our data?
  - particularly in Julia

Dataframe

- Two-dimensional size-mutable, potentially heterogeneous tabular data structure
- Has labeled axes (rows and columns)
  - accessing and plotting data is easier
- Originally introduced in R statistical software

```
## 6×4 DataFrames.DataFrame
##   Row   Student   Height   Gender   Number
##
##   1     1         67.0     female   5
##   2     2         64.0     female   7
##   3     3         61.0     female   2
##   4     4         61.0     female   6
##   5     5         70.0     male     5
##   6     7         61.0     female   3
```

Dataframes vs. Matrices

Dataframe

- has different types of features

```
## 6×2 DataFrames.DataFrame
##   Row   Height   Gender
##
##   1     67.0     female
##   2     64.0     female
##   3     61.0     female
##   4     61.0     female
##   5     70.0     male
##   6     61.0     female
```

Matrix

- all values are of the same type

```
## 4×3 Array{Float64,2}:
##   -0.75    0.7    -0.8
##    0.8    -0.26    1.82
##    1.9     1.86    0.59
##    0.67    0.59    0.48
```

Sample student dataset: studentdata.txt

All students in the introductory statistics course at Bowling Green State University answered the following questions:

1. What is your gender?
2. What is your height in inches?
3. Choose a whole number between 1 and 10.
4. Give the time you went to bed last night.
5. Give the time you woke up this morning.
6. What was the cost (in dollars) of your last haircut, including the tip?
7. Do you prefer water, pop, or milk with your evening meal?

Rich dataset to explore descriptive statistics while illustrating Julia progamming.

Reading data into Julia

We will use the CSV Package to read the tab separated text file. - this returns a Dataframe object

```
#Pkg.add("CSV");
using CSV;
data = CSV.read("studentdata.txt",delim="\t",
                missingstring="NA",rows_for_type_detect=657);
typeof(data)
```

```
## DataFrames.DataFrame
```

Size of a Dataframe

```
size(data)
```

```
## (657, 11)
```

Column names in a Dataframe

```
names(data)
```

```
## 11-element Array{Symbol,1}:
##  :Student
##  :Height
##  :Gender
##  :Shoes
##  :Number
##  :Dvds
##  :ToSleep
##  :WakeUp
##  :Haircut
##  :Job
##  :Drink
```

Columns info. in a Dataframe

```
showcols(data)
```

```
## 11×5 DataFrames.DataFrame. Omitted printing of 2 columns
## Row  variable  eltype                                       nmissing
##
## 1    Student   Int64                                        0
## 2    Height    Float64                                      10
## 3    Gender    CategoricalArrays.CategoricalString{UInt32}  0
## 4    Shoes     Float64                                      22
## 5    Number    Int64                                        2
## 6    Dvds      Float64                                      16
## 7    ToSleep   Float64                                      3
## 8    WakeUp    Float64                                      2
## 9    Haircut   Float64                                      20
## 10   Job       Float64                                      32
## 11   Drink     CategoricalArrays.CategoricalString{UInt32}  11
```

Head of a Dataframe

```
# first 6 rows in Dataframe
head(data)
```

```
## 6×11 DataFrames.DataFrame. Omitted printing of 3 columns
##  Row  Student  Height  Gender  Shoes    Number  Dvds  ToSlee
##
##  1    1        67.0    female  10.0     5       10.0  -2.5
##  2    2        64.0    female  20.0     7       5.0   1.5
##  3    3        61.0    female  12.0     2       6.0   -1.5
##  4    4        61.0    female  3.0      6       40.0  2.0
##  5    5        70.0    male    4.0      5       6.0   0.0
##  6    6        63.0    female  missing  3       5.0   1.0
```

Tail of Dataframe

```
# bottom 6 rows in Dataframe
tail(data)
```

```
## 6×11 DataFrames.DataFrame. Omitted printing of 3 columns
##  Row  Student  Height  Gender  Shoes  Number  Dvds  ToSleep
##
##  1    652      68.0    female  30.0   6       4.0   0.5
##  2    653      71.0    female  15.0   8       25.0  -1.0
##  3    654      66.0    female  25.0   5       1.0   -1.5
##  4    655      67.0    female  10.0   7       10.0  0.0
##  5    656      68.0    male    3.0    5       15.0  2.5
##  6    657      69.0    male    4.0    6       20.0  3.5
```

Drop rows with missing values

```
size(data)
```

```
## (657, 11)
```

```
data = dropmissing(data);
size(data)
```

```
## (559, 11)
```

Summary statistics of a column

```
describe(data[:,[:Height]])
```

```
## 1×8 DataFrames.DataFrame. Omitted printing of 1 columns
##   Row   variable   mean      min    median   max    nunique   nmissing
##
##   1     Height     66.7648   54.0   67.0     84.0             0
```

Summary statistics for separate groups

```
describe(data[find(data[:Gender].=="male"),[:Height]])
```

```
## 1×8 DataFrames.DataFrame. Omitted printing of 1 columns
##   Row   variable   mean      min    median   max    nunique   nmissi
##
##   1     Height     70.3782   59.0   71.0     79.0             0
```

```
describe(data[find(data[:Gender].=="female"),[:Height]])
```

```
## 1×8 DataFrames.DataFrame. Omitted printing of 1 columns
##   Row   variable   mean     min    median   max    nunique   nmissin
##
##   1     Height     64.829   54.0   65.0     84.0             0
```

Unique values in a column

```
unique(data[:Drink])
```

```
## 3-element Array{Union{Missings.Missing, String},1}:
##  "water"
##  "pop"
##  "milk"
```

Grouping rows based on a column

```
# returns GroupedDataFrame - vector of Dataframes
ans = groupby(data, :Drink)
```

```
## DataFrames.GroupedDataFrame  3 groups with keys: Symbol[:Drink]
## First Group:
## 308×11 DataFrames.SubDataFrame{Array{Int64,1}}. Omitted printing of 3 co
##   Row   Student   Height   Gender   Shoes   Number   Dvds   ToSleep   WakeUp
##
##   1     1         67.0     female   10.0    5        10.0   -2.5      5.5
##   2     4         61.0     female   3.0     6        40.0   2.0       8.5
##   3     7         61.0     female   12.0    3        53.0   1.5       7.5
##   4     9         66.0     female   30.0    3        40.0   -0.5      7.0
##   5     12        63.0     female   20.0    4        60.0   -1.0      7.0
##   6     16        65.0     female   40.0    7        50.0   -0.5      7.0
##   7     19        71.0     male     6.0     7        0.0    0.5       7.5
##   8     20        64.0     female   4.0     7        8.0    -1.5      7.5
##
##   300   638       67.0     female   12.0    8        13.0   4.0       9.0
##   301   639       66.5     female   13.0    5        48.0   2.5       9.0
##   302   641       63.0     female   10.0    3        6.0    1.5       11.0
##   303   640       66.0     female   24.0    3        12.0   -1.0      8.0
```

Finding number of rows for each type

```
# apply nrow function to every row in the group from Drink
ans = by(data, :Drink, nrow)
```

```
## 3×2 DataFrames.DataFrame
##   Row   Drink   x1
##
##   1     water   308
##   2     pop     154
##   3     milk    97
```

Adding a new column to DataFrame

```
# apply nrow function to every row in the group from Drink
data[:HrsSleep] = data[:WakeUp] - data[:ToSleep];
names(data)
```

```
## 12-element Array{Symbol,1}:
##  :Student
##  :Height
##  :Gender
##  :Shoes
##  :Number
##  :Dvds
##  :ToSleep
##  :WakeUp
##  :Haircut
##  :Job
##  :Drink
##  :HrsSleep
```

Creating a Dataframe from an existing Dataframe

```
data1 = data[:,[:Student,:Height,:Gender,:Number]]
```

```
## 559×4 DataFrames.DataFrame
##   Row   Student   Height   Gender   Number
##
##   1     1         67.0     female   5
##   2     2         64.0     female   7
##   3     3         61.0     female   2
##   4     4         61.0     female   6
##   5     5         70.0     male     5
##   6     7         61.0     female   3
##   7     8         64.0     female   4
##   8     9         66.0     female   3
##
##   551   648       68.0     female   7
##   552   649       65.0     female   5
##   553   650       74.0     male     2
##   554   651       72.0     male     7
##   555   652       68.0     female   6
```

Creating a Dataframe from vectors

```
df = DataFrame(x=1:10, y=rand(10), label="a");
head(df)
```

```
## 6×3 DataFrames.DataFrame
##   Row   x   y          label
##
##   1     1   0.420658   a
##   2     2   0.857426   a
##   3     3   0.683784   a
##   4     4   0.352729   a
##   5     5   0.131234   a
##   6     6   0.258234   a
```

Creating a Dataframe form a matrix

```
x = rand(3, 4)
```

```
## 3×4 Array{Float64,2}:
##  0.129378  0.855021  0.55745   0.492016
##  0.493085  0.222304  0.308584  0.854161
##  0.586671  0.341275  0.846363  0.679096
```

```
df = convert(DataFrame, x);
head(df)
```

```
## 3×4 DataFrames.DataFrame
##   Row   x1         x2         x3         x4
##
##   1     0.129378   0.855021   0.55745    0.492016
##   2     0.493085   0.222304   0.308584   0.854161
##   3     0.586671   0.341275   0.846363   0.679096
```

Vertical concatenation of Dataframes

```
df1 = DataFrame(x=1:5, y=rand(5), label="a");
df2 = DataFrame(x=1:5, y=rand(5), label="b");
df = vcat(df1, df2)
```

```
## 10×3 DataFrames.DataFrame
##   Row   x   y            label
##
##   1     1   0.941676     a
##   2     2   0.768907     a
##   3     3   0.742488     a
##   4     4   0.618849     a
##   5     5   0.00647757   a
##   6     1   0.2857       b
##   7     2   0.106403     b
##   8     3   0.950561     b
##   9     4   0.182568     b
##   10    5   0.433047     b
```

Horizontal concatenation of Dataframes

```
df1 = DataFrame(w=rand(5), x=rand(5));
df2 = DataFrame(y=rand(5), z=rand(5));
df = hcat(df1, df2)
```

```
## 5×4 DataFrames.DataFrame
##   Row   w          x          y           z
##
##   1     0.117431   0.556622   0.241158    0.349199
##   2     0.271952   0.78754    0.635181    0.934288
##   3     0.741677   0.897741   0.603863    0.922777
##   4     0.430203   0.221894   0.171996    0.866495
##   5     0.268923   0.301041   0.0581873   0.813511
```