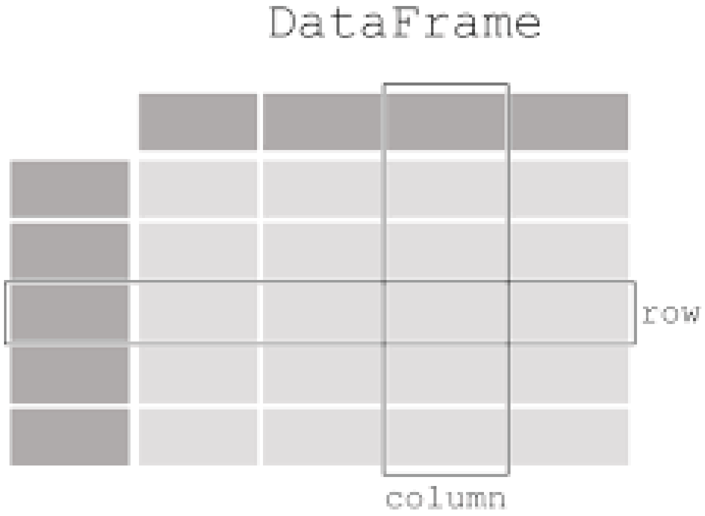
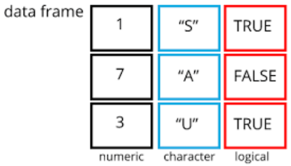
Lesson 02
Data Structures
Data Frames

Outline

- Data Frames
  - Construction
  - Basic attributes
  - Rows and columns
  - Factor columns





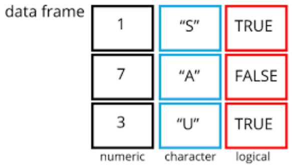
data frame

numeric	character	logical
1	"S"	TRUE
7	"A"	FALSE
3	"U"	TRUE

Data Frames – Definition

- A rectangular data structures
  - Columns representing variables (attributes, characteristics)
    - Interest rates, loan type, …
    - Can be of different data types
    - Vectors of the same length
  - Rows represent observations
    - Customer’s loan with specific interest rate and loan type
- Similar to spreadsheets
  - Better comparison are relational database tables
- Creating data frames
  - As a collection of existing vectors
  - By reading data from a file or a database

Data Frames - Construction



- Using `data.frame` function
  - Open **Struct1_DataFrames.r**
  - Review the existing vectors on characteristics of the ten loans

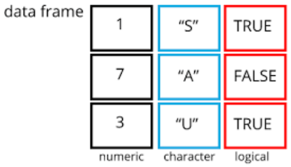
```
# Create a data frame using data.frame function on the provided vectors
loans_df <- data.frame(amount, intRate, loanTerm, loanType, mthPmt)
loans_df
```

  - Give columns appropriate names

```
# Create a vector of column names
loan_cols <- c("Amount", "Rate", "Term", "Type", "Payment")
names(loans_df) <- loan_cols
loans_df
```

- Using the same data stored in a CSV file
  - See next module on reading data into R

Data Frames – Basic Attributes



data frame

1	“S”	TRUE
7	“A”	FALSE
3	“U”	TRUE
numeric	character	logical

- Determine the size of the data frame

```
nrow(loans_df)
ncol(loans_df)
dim(loans_df)
```

- Get the structure of the data frame

```
str(loans_df)
```

- Display some and all column names

```
names(loans_df)
names(loans_df)[c(2,4)]
```

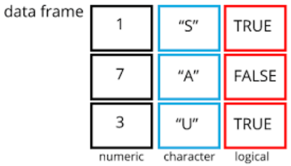
- Show only the top and bottom portions of a large data frame

```
head(loans_df)
tail(loans_df, n=1)
```

- Print the basic summaries of the data in a data frame

```
summary(loans_df)
```

Data Frames – Rows and Columns



data frame

numeric	character	logical
1	"S"	TRUE
7	"A"	FALSE
3	"U"	TRUE

- Accessing individual columns of a data frame

```
loans_df$Type
loans_df[2]
loans_df[,1]
```

- You constantly have to be vigilant about data types

  `class(loans_df$Type)` returns `“character”` vector

  `is.vector(loans_df$Type)` returns `TRUE`

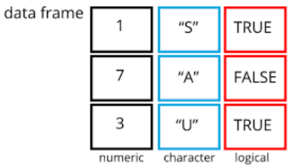
  `class(loans_df[2])` returns `“data.frame”`

  `is.data.frame(loans_df[2])` returns `TRUE`

- Various ways of accessing subsets of rows and columns
  - Need to experiment on your own, be mindful of the resulting data type

```
# Accessing consecutive rows and a single column
loans_df[4:7,2]
# Accessing nonconsecutive rows and all columns
loans_df[c(1,3,5),]  # Experiment with removing the comma
# Accessing columns using their names
loans_df[,c("Rate","Type")] # Experiment with removing the comma
```

Data Frames – Factor Columns



numeric	character	logical
1	"S"	TRUE
7	"A"	FALSE
3	"U"	TRUE

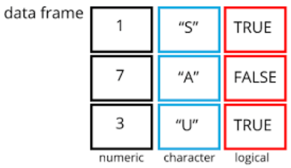
- Type is a `character`, rather than `factor` column

```
loans_df["Type"]
class(loans_df["Type"]) # data.frame column
loans_df[["Type"]] # Double-brackets more used with lists
class(loans_df[["Type"]]) # character vector
# Paramter drop=FALSE assures data frame type
loans_df[,"Type", drop=FALSE]
class(loans_df[,"Type", drop=FALSE])
```

- Recreating the same data frame with Type as factor

```
loanTypeFactor <- as.factor(loanType)
loans_df2 <- data.frame(amount, intRate, loanTerm,
loanTypeFactor, mthPmt)
names(loans_df2) <- loan_cols
class(loans_df2[, "Type"])
```

Data Frames – Indicator Variables



data frame

1	“S”	TRUE
7	“A”	FALSE
3	“U”	TRUE
numeric	character	logical

- Indicator (dummy) variables
  - Critical for analysis using categorical variables
  - Gender variable: 1=Male; 0=Female
  - Regression of salary on various variables, including gender
  - The regression coefficient is the salary difference between male and female employees
  - Allows to determine significance, etc..
- Creating dummy variables for loan type

```
model.matrix(~loanTypeFactor - 1)
```

Summary

- Examined `data.frame` data structure
  - Rectangular structure similar to DB tables
  - Columns are attributes, rows observations
  - One of the most widely used data structures in data science
- Discussed main data frame concepts
  - Construction, attributes: size, names, head, tail, summary
  - Row and column access and basic sub-setting operations
  - Converting categorical columns into factors for subsequent statistical analyses