INTRODUCTION TO R AND RSTUDIO

Part 2: `tidyverse` (follow along in RStudio)

https://econ.lange-analytics.com/aibook/

LEARNING OUTCOMES

What you will learn in this session:

- The Structure of R commands
- About the `tidyverse` package for data frames
  - `select()` and rename columns (variables)
  - `filter()` rows (observations)
  - `mutate()` (define columns (variables); overwrite old or create new)
  - piping (connecting commands) with `%>%`.

BASICS OF R COMMANDS

R commands consists of the **command's name followed by a pair of parentheses**: `command()`

Inside the `()` we can define one or more **arguments** for the command.

```r
VecTest=c(1,2,3)
```

```r
cat("Sum:", sum(x=VecTest))
```

```
Sum: 6
```

```r
cat("Mean:", mean(VecTest))
```

```
Mean: 2
```

- Arguments in a command usually have names such as `x=` or `data=`
- R does not require to use the argument’s name, but **order matters**
- R commands have many arguments. Most have default values
- We can nest commands. However, nesting too deeply makes code difficult to read.

:::

STRUCTURE OF R COMMANDS

Most R commands have the following structure:

$$\underbrace{DataNew}_{\text{R object storing the result}} = \underbrace{Command}_{\text{Name of the command}} \underbrace{(\overbrace{Data}^{\text{1. Argument: Data to process}}, \overbrace{Arg2, Arg3, \ldots, ArgN}^{\text{More Arguments}})}_{\text{Arguments inside () and separated by komma}}$$

Often the `data` argument is the first argument in a command. Usually named `data=` or `x=`.

USE A COMMAND WITH AND WITHOUT ARGUMENT NAMES 🤓

```r
VecTest=c(1,2,3)
```

```r
Result=mean(x=VecTest, trim=0, na.rm=FALSE)
cat("The mean of the values in vector VecTest is:", Result)
```

```
The mean of the values in vector VecTest is: 2
```

```r
Result=mean(VecTest, 0, FALSE)
cat("The mean of the values in vector VecTest is:", Result)
```

```
The mean of the values in vector VecTest is: 2
```

```r
Result=mean(VecTest)
cat("The mean of the values in vector VecTest is:", Result)
```

```
The mean of the values in vector VecTest is: 2
```

All three examples are equivalent

Try `? mean` in the Rstudio console to see the default values.

IMPORTANT COMANDS FROM `tidyverse/dplyr` PACKAGE

- `dplyr` package is part of the `tidyverse` (meta) package
- `library(tidyverse)` (loads the `tidyverse` and its packages)
- `select()` selects columns (variables) from a data frame
- `filter()` filters rows (observations) for specific criteria
- `mutate()` calculates new or overwrites existing columns (variables) based on other columns (just like Excel).

TITANIC DATASET

```r
library(rio)
DataTitanic=import("https://lange-analytics.com/AIBook/Data/TitanicDataCl.csv")
head(DataTitanic)
```

```
  Survived Pclass                                               Name    Sex Age
1        0      3                             Mr. Owen Harris Braund   male  22
2        1      1 Mrs. John Bradley (Florence Briggs Thayer) Cumings female  38
3        1      3                              Miss. Laina Heikkinen female  26
4        1      1        Mrs. Jacques Heath (Lily May Peel) Futrelle female  35
5        0      3                            Mr. William Henry Allen   male  35
6        0      3                                    Mr. James Moran   male  27
  SiblingsSpousesAboard ParentsChildrenAboard FareInPounds
1                     1                     0       7.2500
2                     1                     0      71.2833
3                     0                     0       7.9250
4                     1                     0      53.1000
5                     0                     0       8.0500
6                     0                     0       8.4583
```

THE select() COMMAND

- select(DataMine, Var1, Var2) selects columns (variables) Var1 and Var2 from a data frame DataMine. The first argument is the data= argument followed by the names of the selected variables.
- select(Data, -Var1, -Var2) selects all columns (variables) except Var1 and Var2 from a data frame DataMine.

Here is an example using the DataTitanic data frame from the previous slide:

```r
library(tidyverse)
DataTitanicSelVar=select(DataTitanic,Survived, Name, Sex, Age)
head(DataTitanicSelVar)
```

```
  Survived                                               Name    Sex Age
1        0                             Mr. Owen Harris Braund   male  22
2        1 Mrs. John Bradley (Florence Briggs Thayer) Cumings female  38
3        1                              Miss. Laina Heikkinen female  26
```

THE filter() COMMAND

The `filter()` command filters rows (observations) of a data frame for specific criteria. The first argument is the `data=` argument followed by the filter criteria.

E.g., *filter* for female passengers from the dataset: Use `DataTitanicSelVar` that we created in the previous slide (note that we have to use `==` instead of `=` for the criteria):

```r
DataTitanicSelVarFem=filter(DataTitanicSelVar, Sex=="female")
head(DataTitanicSelVarFem)
```

```
  Survived                                               Name    Sex Age
1        1 Mrs. John Bradley (Florence Briggs Thayer) Cumings female  38
2        1                              Miss. Laina Heikkinen female  26
3        1        Mrs. Jacques Heath (Lily May Peel) Futrelle female  35
4        1   Mrs. Oscar W (Elisabeth Vilhelmina Berg) Johnson female  27
5        1                 Mrs. Nicholas (Adele Achem) Nasser female  14
6        1                     Miss. Marguerite Rut Sandstrom female   4
```

THE mutate() COMMAND 🤓

mutate() creates or overwrites columns (variables) based on other columns (just like Excel). The first argument is the data= argument followed by the instructions on how to create the new variable.

E.g., *mutate* calculates new column Born based on Age during Titanic disaster (1912). Uses DataTitanicSelVarFem from previous slide:

```
DataTitatincSelVarFemBirthYear=mutate(DataTitanicSelVarFem, Born=1912-Age)
head(DataTitatincSelVarFemBirthYear)
```

```
  Survived                                                Name    Sex Age Born
1        1 Mrs. John Bradley (Florence Briggs Thayer) Cumings female  38 1874
2        1                              Miss. Laina Heikkinen female  26 1886
3        1        Mrs. Jacques Heath (Lily May Peel) Futrelle female  35 1877
4        1   Mrs. Oscar W (Elisabeth Vilhelmina Berg) Johnson female  27 1885
5        1                 Mrs. Nicholas (Adele Achem) Nasser female  14 1898
6        1                    Miss. Marguerite Rut Sandstrom female   4 1908
```

SUMMARY

1. We selected variables $Survived$, $Name$, Sex, Age and saved in `DataTitanicSelVar`
2. We filtered for females and saved in `DataTitanicSelVarFem`
3. We mutated to calculate new variable and saved finally in `DataTitanicSelVarFemBirthYear`

Could this be done easier?

Note, overwriting data frames such as `DataTitanic` is usually a bad idea!

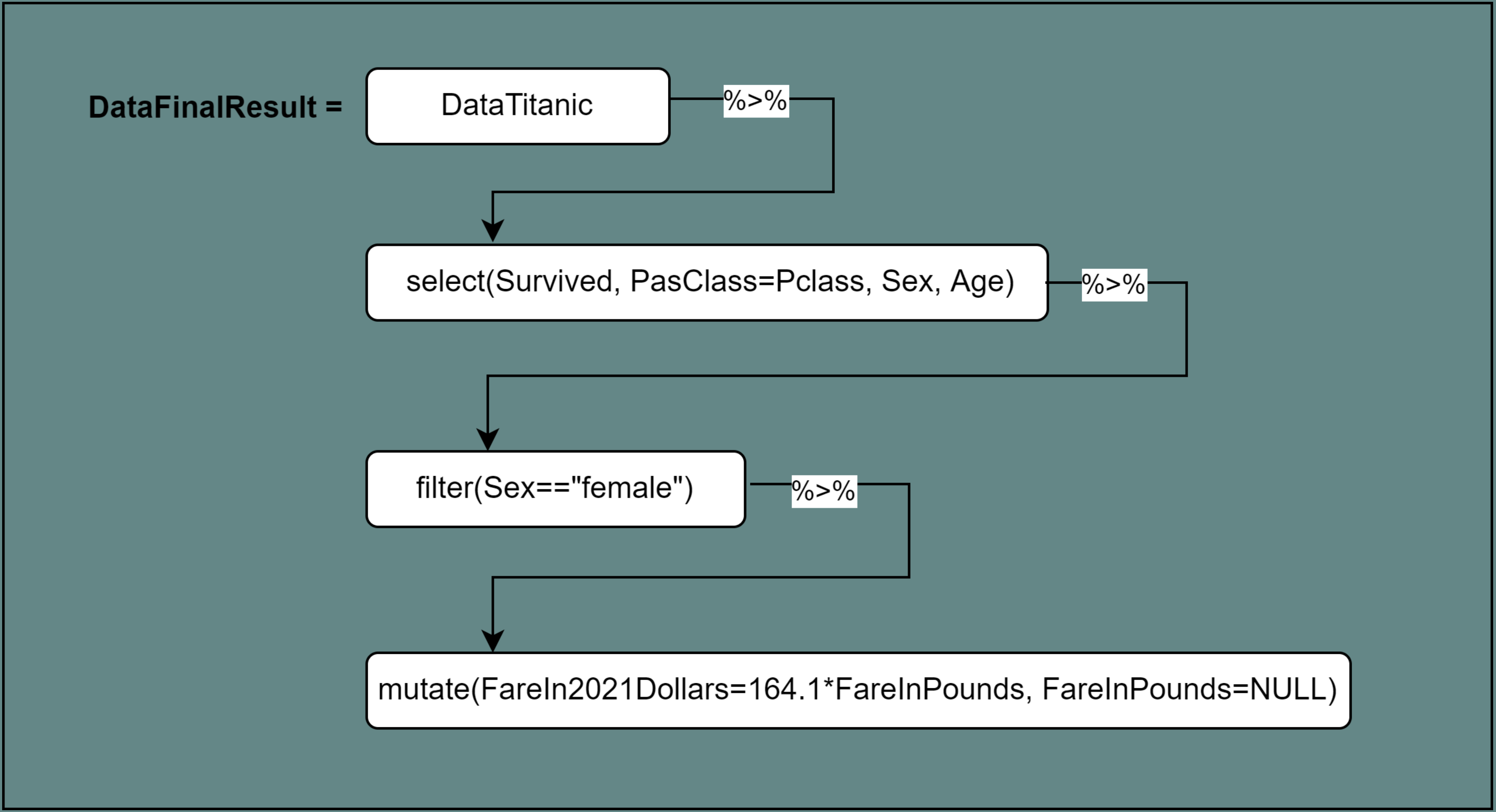
ALTERNATIVE: NESTING

(I AM NOT SERIOUS)

```r
library(tidyverse)
DataTitanicFinal= mutate(
        filter(select(DataTitanic,Survived, Name, Sex, Age),
               Sex=="female"),
                        Born=1912-Age)
head(DataTitanicFinal)
```

```
  Survived                                               Name    Sex Age Born
1        1 Mrs. John Bradley (Florence Briggs Thayer) Cumings female  38 1874
2        1                              Miss. Laina Heikkinen female  26 1886
3        1        Mrs. Jacques Heath (Lily May Peel) Futrelle female  35 1877
4        1   Mrs. Oscar W (Elisabeth Vilhelmina Berg) Johnson female  27 1885
5        1                 Mrs. Nicholas (Adele Achem) Nasser female  14 1898
6        1                     Miss. Marguerite Rut Sandstrom female   4 1908
```

PIPING SCHEMA



```
DataFinalResult = DataTitanic %>%
select(Survived, PasClass=Pclass, Sex, Age) %>%
filter(Sex=="female") %>%
mutate(FareIn2021Dollars=164.1*FareInPounds, FareInPounds=NULL)
```

Piping Schema

ALTERNATIVE: PIPING

(WILL BE USED THROUGHOUT THE COURSE/BOOK) 🤓

```r
library(tidyverse)
DataTitanicFinal= DataTitanic %>%
                  select(Survived, Name, Sex, Age) %>%
                  filter(Sex=="female") %>%
                  mutate(Born=1912-Age)
head(DataTitanicFinal)
```

```
  Survived                                               Name    Sex Age Born
1        1 Mrs. John Bradley (Florence Briggs Thayer) Cumings female  38 1874
2        1                              Miss. Laina Heikkinen female  26 1886
3        1        Mrs. Jacques Heath (Lily May Peel) Futrelle female  35 1877
4        1   Mrs. Oscar W (Elisabeth Vilhelmina Berg) Johnson female  27 1885
5        1                 Mrs. Nicholas (Adele Achem) Nasser female  14 1898
6        1                     Miss. Marguerite Rut Sandstrom female   4 1908
```

https://econ.lange-analytics.com/aibook/

QUESTIONS