INTRODUCTION TO R AND RSTUDIO

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

https://ai.lange-analytics.com/book/

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.

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

```
1 sum(x=VecTest)
```

```
[1] 6
```

```
1 mean(VecTest)
```

```
[1] 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 

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

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

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

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

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

```
1  Result=mean(VecTest)
2  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 COMMANDS 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

```
1 library(rio)
2 DataTitanic=import("https://ai.lange-analytics.com/data/Titanic.csv")
3 head(DataTitanic)
```

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

```
1 library(tidyverse)
2 DataTitanicSelVar=select(DataTitanic,Survived, Name, Sex, Age)
3 head(DataTitanicSelVar)
```

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

```
4         1 Mrs. Jacques Heath Futrelle female  35
5         0     Mr. William Henry Allen   male  35
6         0             Mr. James Moran   male  27
```

```
»
```

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):

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

```
  Survived                                Name    Sex Age
1        1           Mrs. John Bradley Cumings female  38
2        1               Miss. Laina Heikkinen female  26
3        1         Mrs. Jacques Heath Futrelle female  35
4        1                Mrs. Oscar W 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:

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

```
  Survived                                Name    Sex Age Born
1        1          Mrs. John Bradley Cumings female  38 1874
2        1              Miss. Laina Heikkinen female  26 1886
3        1        Mrs. Jacques Heath Futrelle female  35 1877
4        1             Mrs. Oscar W 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

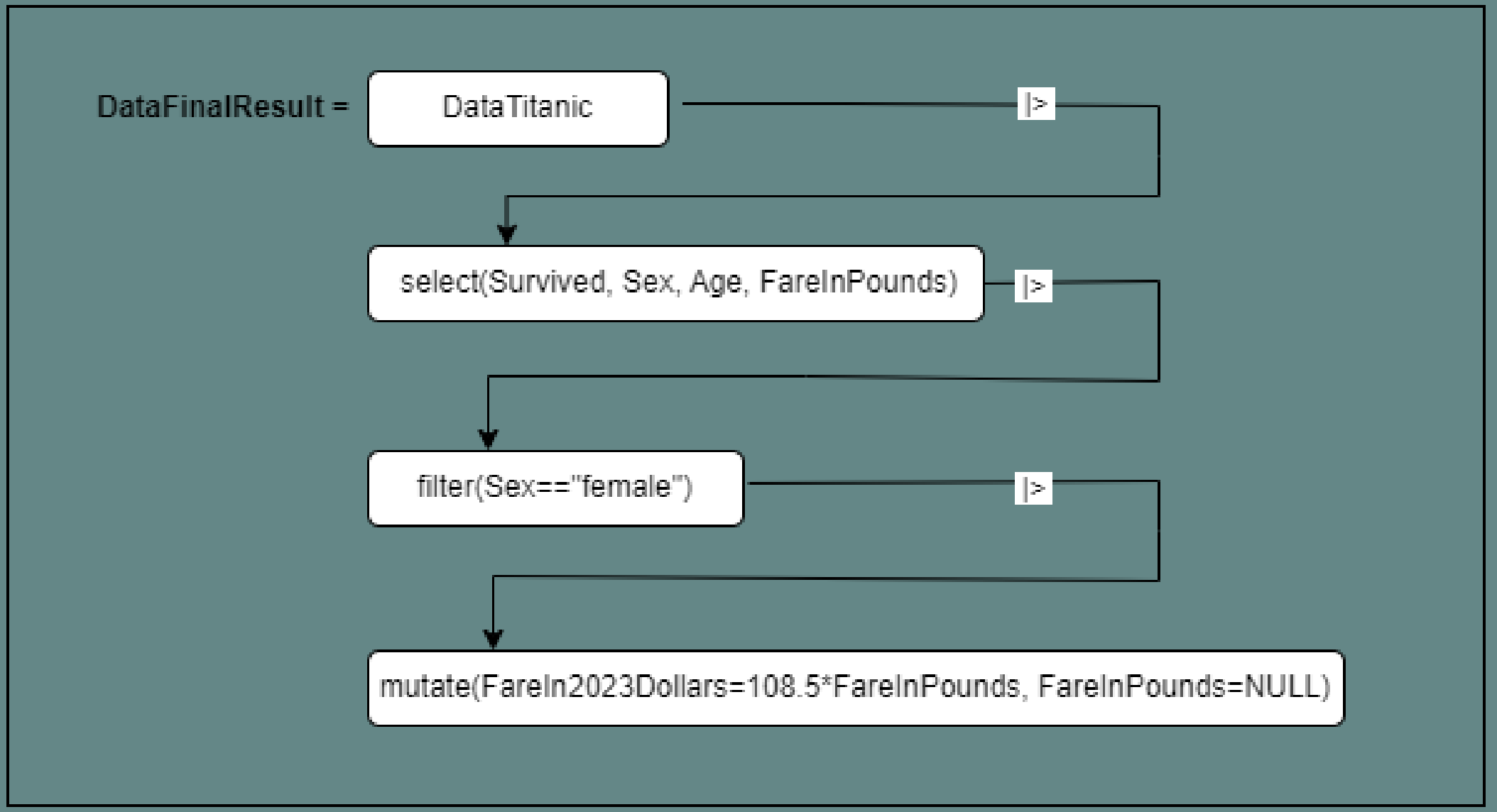
(I AM NOT SERIOUS)

```
1 library(tidyverse)
2 DataTitanicFinal= mutate(
3         filter(select(DataTitanic,Survived, Name, Sex, Age),
4                Sex=="female"),
5                         Born=1912-Age)
6 head(DataTitanicFinal)
```

```
  Survived                                Name    Sex Age Born
1        1          Mrs. John Bradley Cumings female  38 1874
2        1              Miss. Laina Heikkinen female  26 1886
3        1        Mrs. Jacques Heath Futrelle female  35 1877
4        1             Mrs. Oscar W 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



Piping Schema

ALTERNATIVE: PIPING

(WILL BE USED THROUGHOUT THE COURSE/BOOK) 🤓

```
1 library(tidyverse)
2 DataTitanicFinal= DataTitanic |>
3                   select(Survived, Name, Sex, Age) |>
4                   filter(Sex=="female") |>
5                   mutate(Born=1912-Age)
6 head(DataTitanicFinal)
```

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

»

https://ai.lange-analytics.com/book/

QUESTIONS

https://ai.lange-analytics.com/book/