Introduction to Statistics and R

Accessing and managing subsets of data

Eric Stemmler

Khovd University

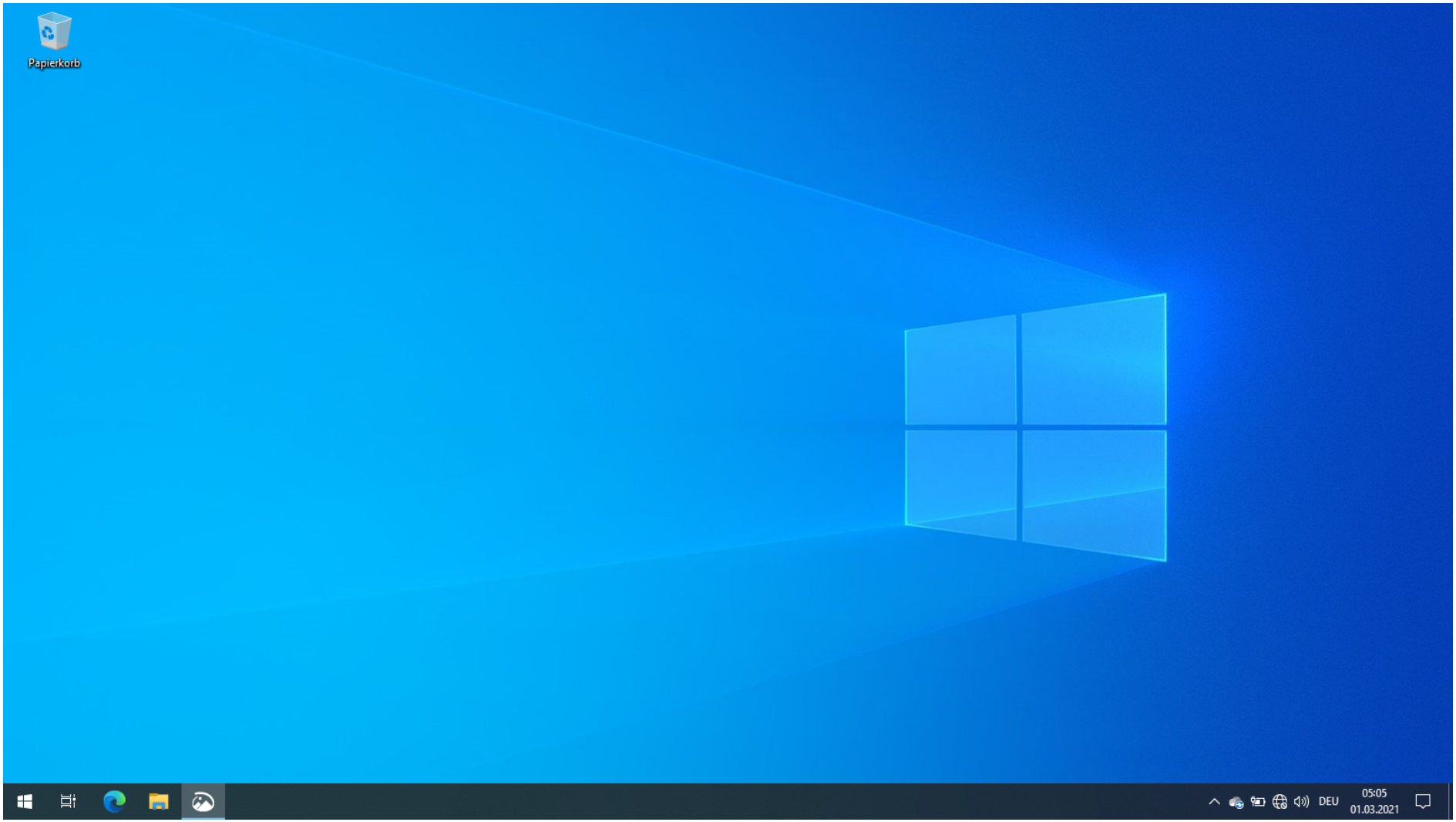
03.03.2021

8 Quiz

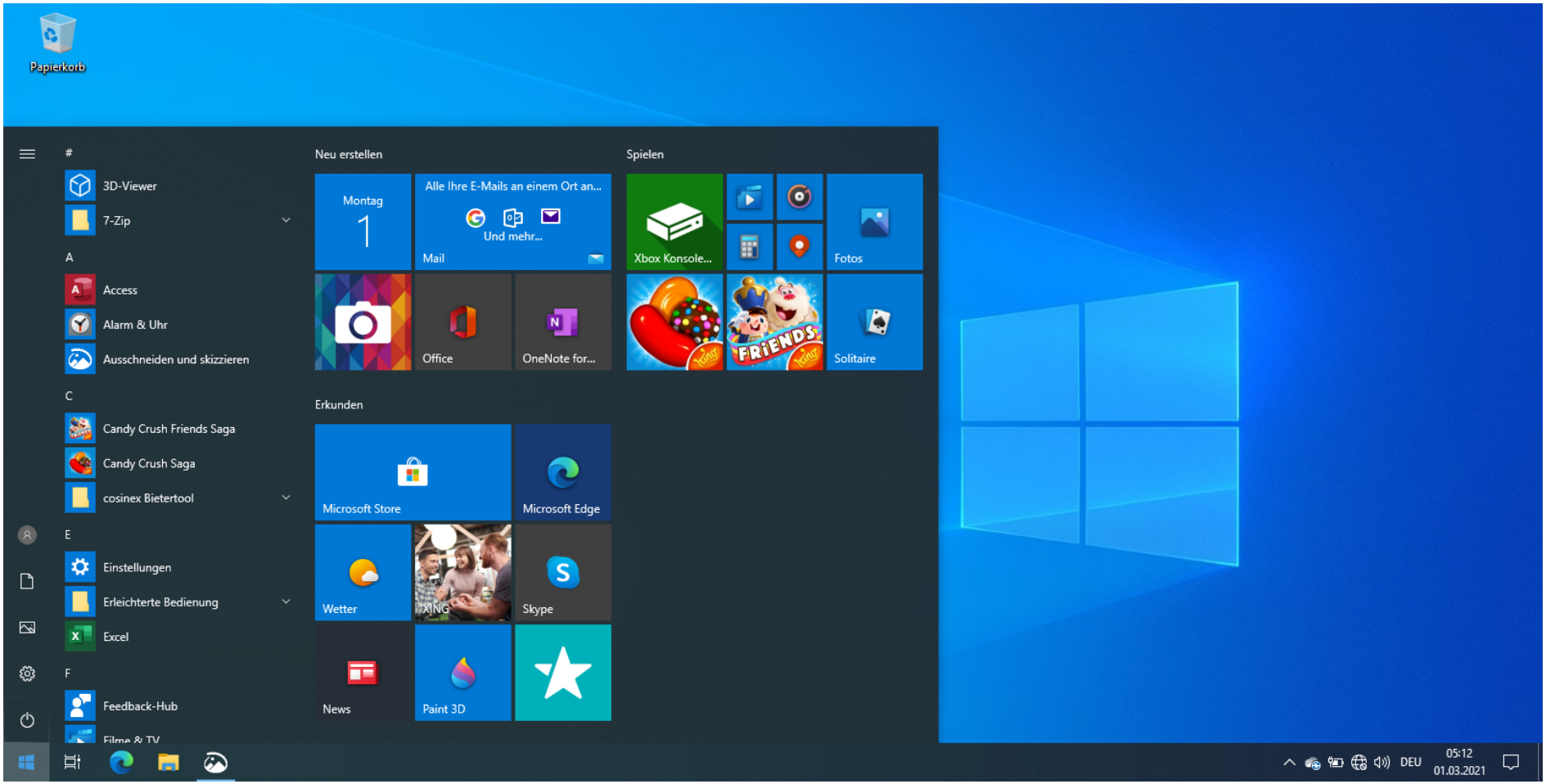
Section 1

Recap: Using RGui

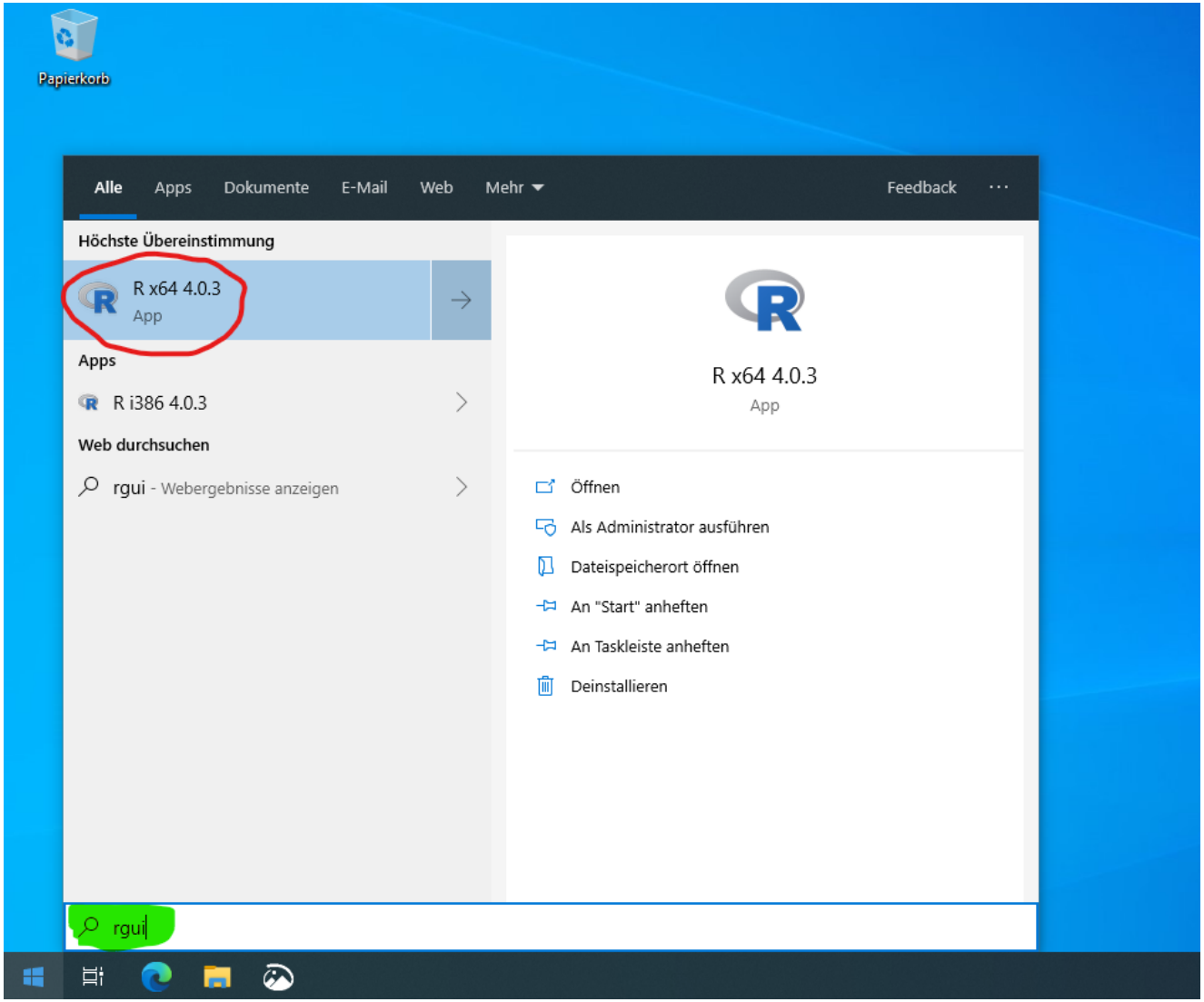
Recap: Using RGui



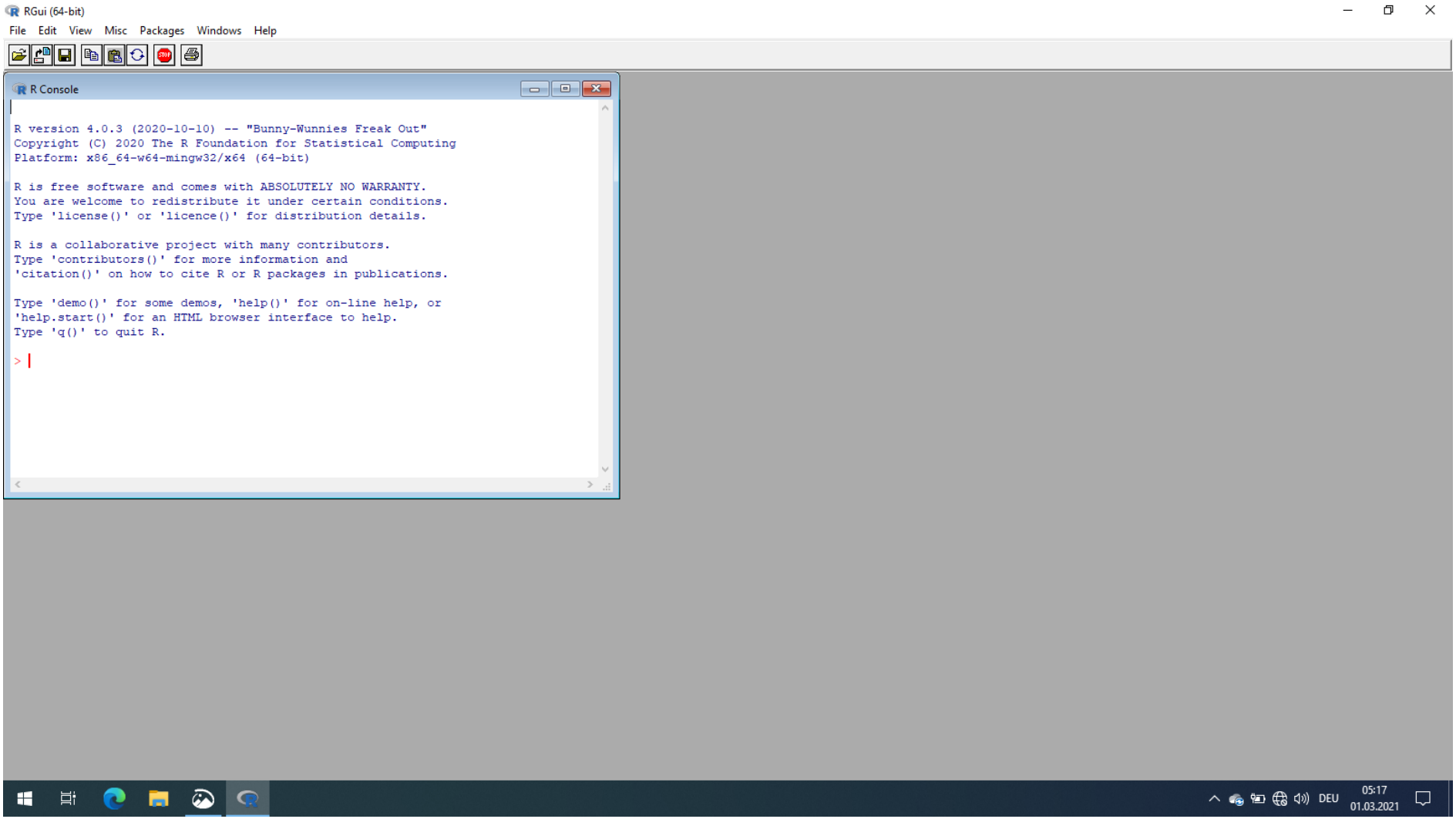
Recap: Using RGui



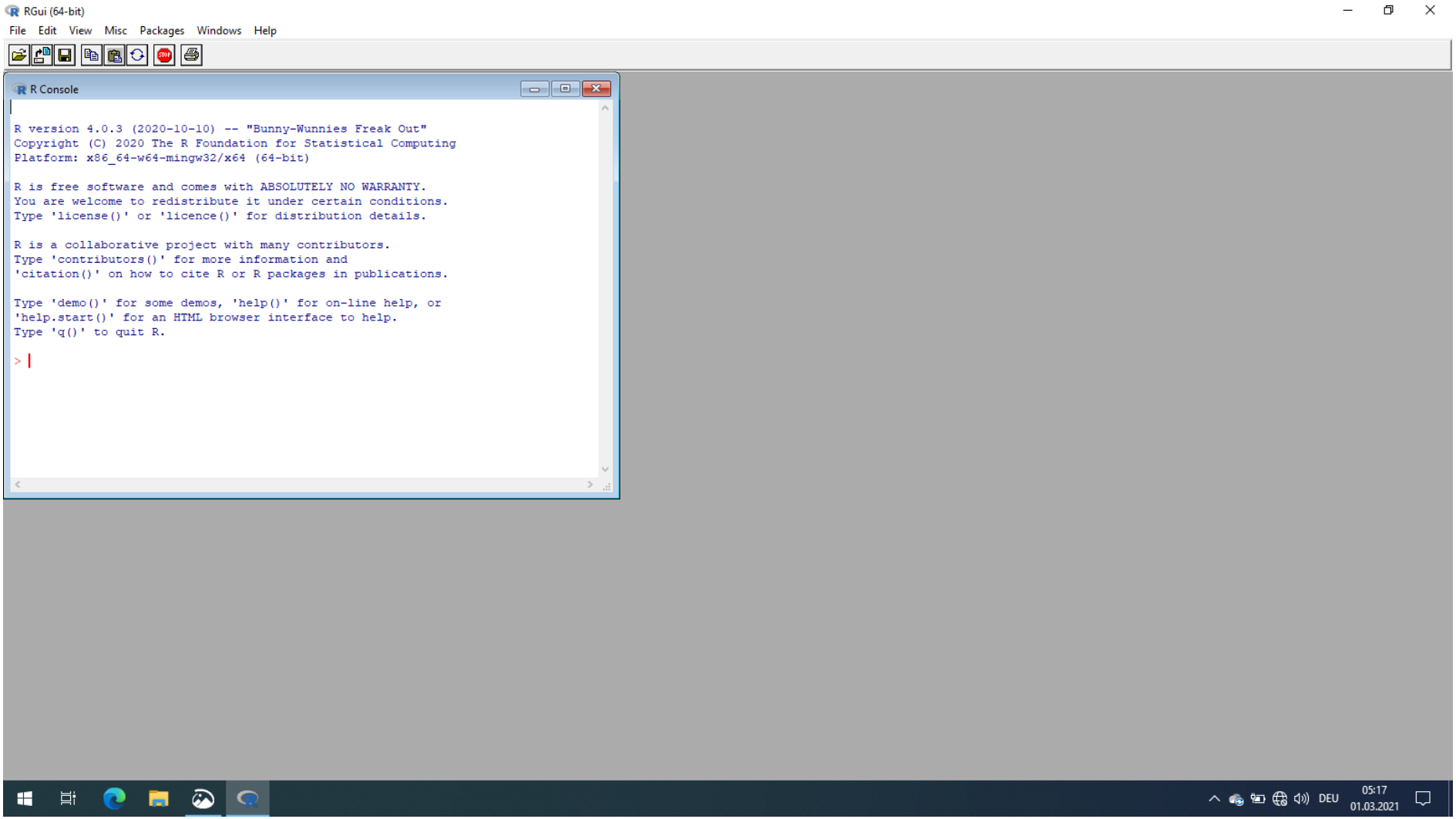
Recap: Using RGui



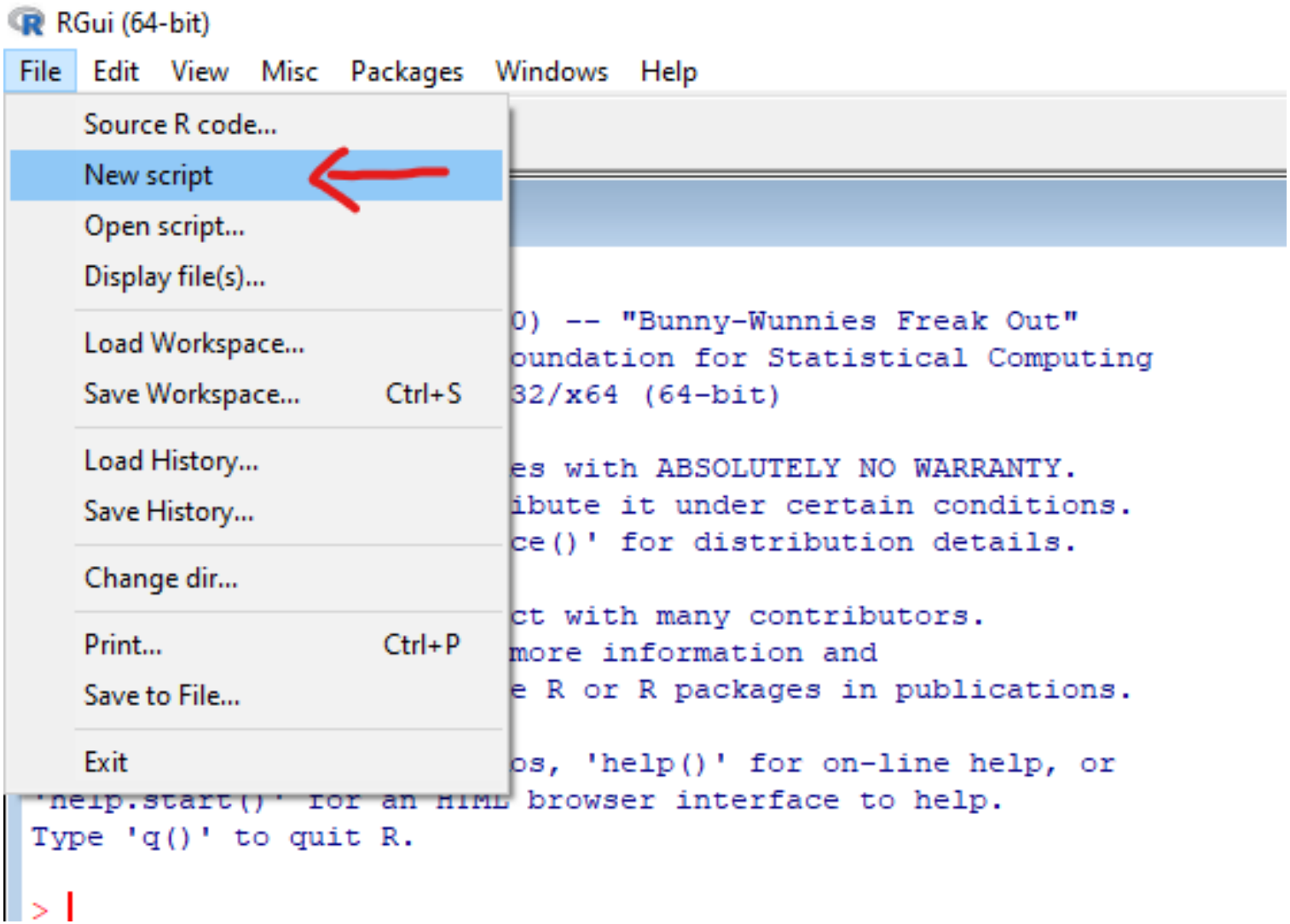
Recap: Using RGui



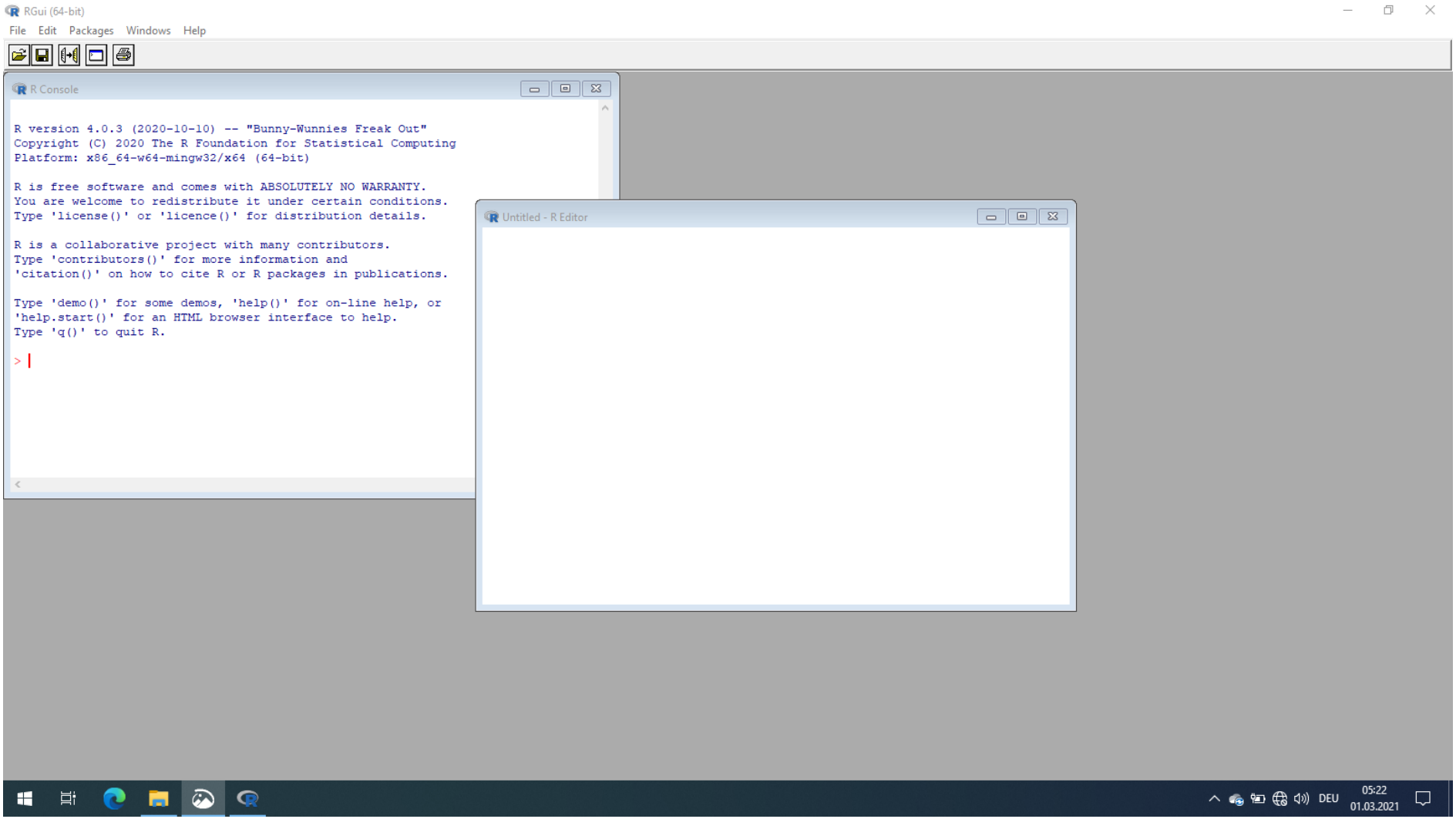
Recap: Using RGui



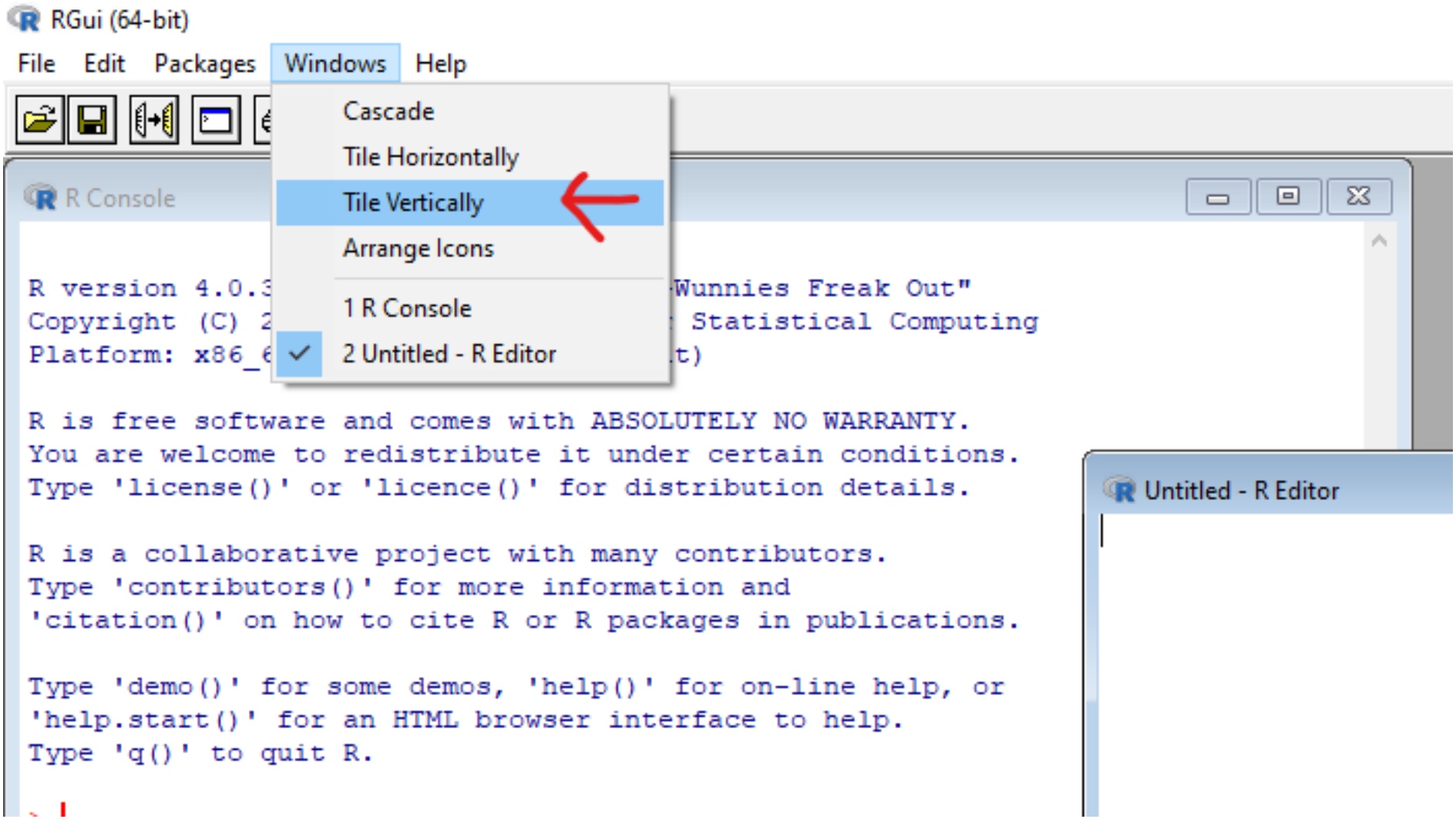
Recap: Using RGui



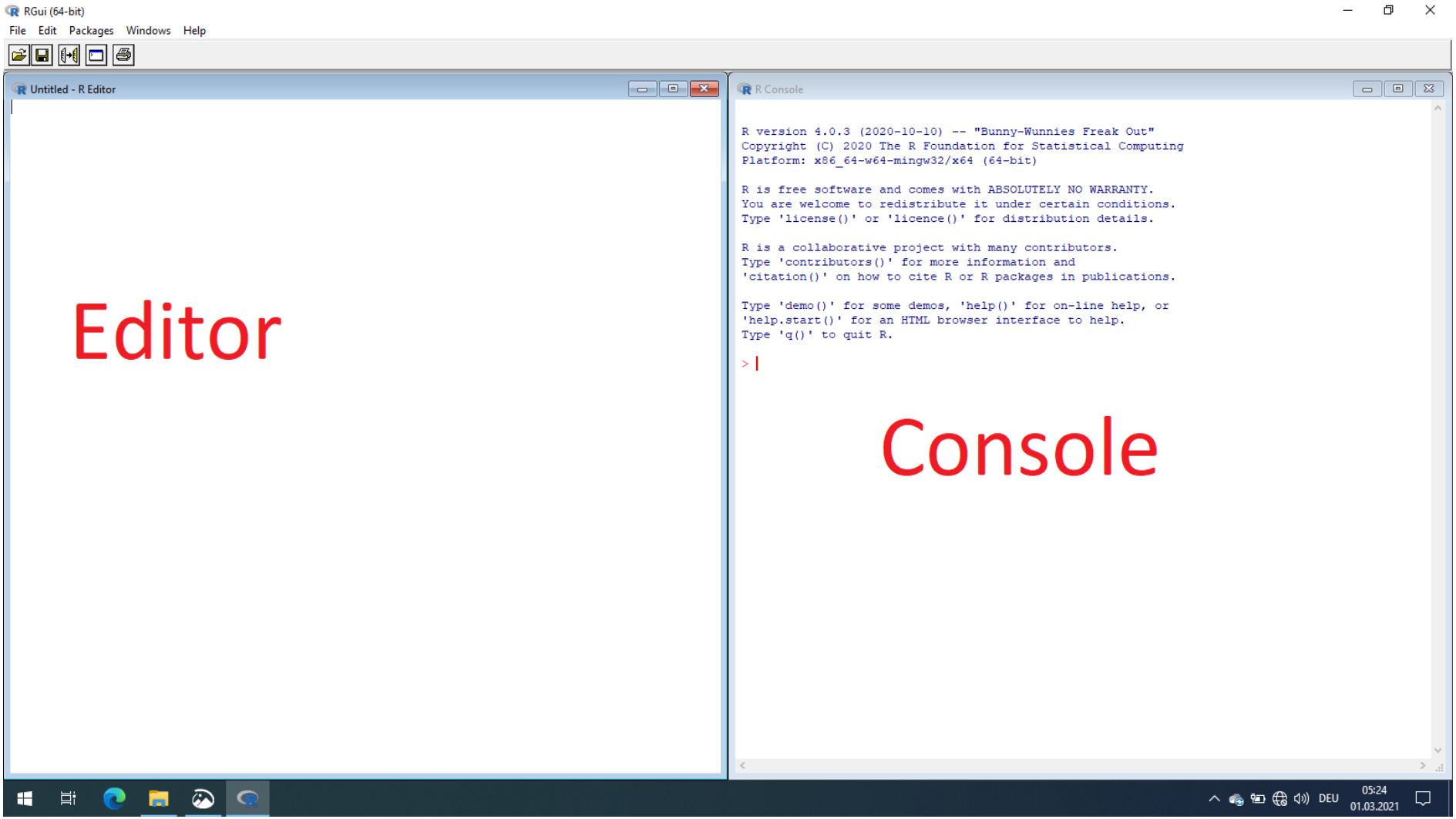
Recap: Using RGui



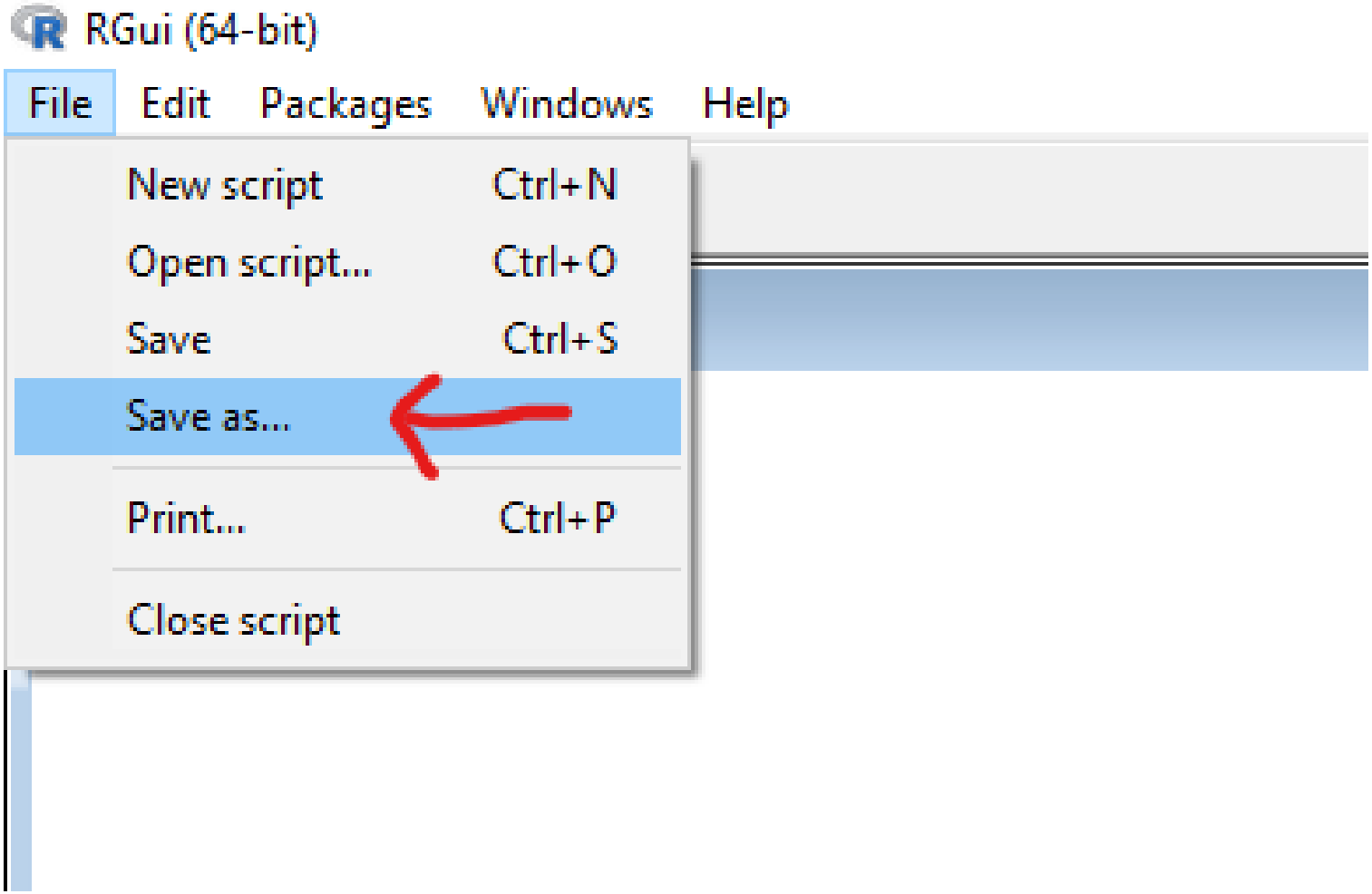
Recap: Using RGui



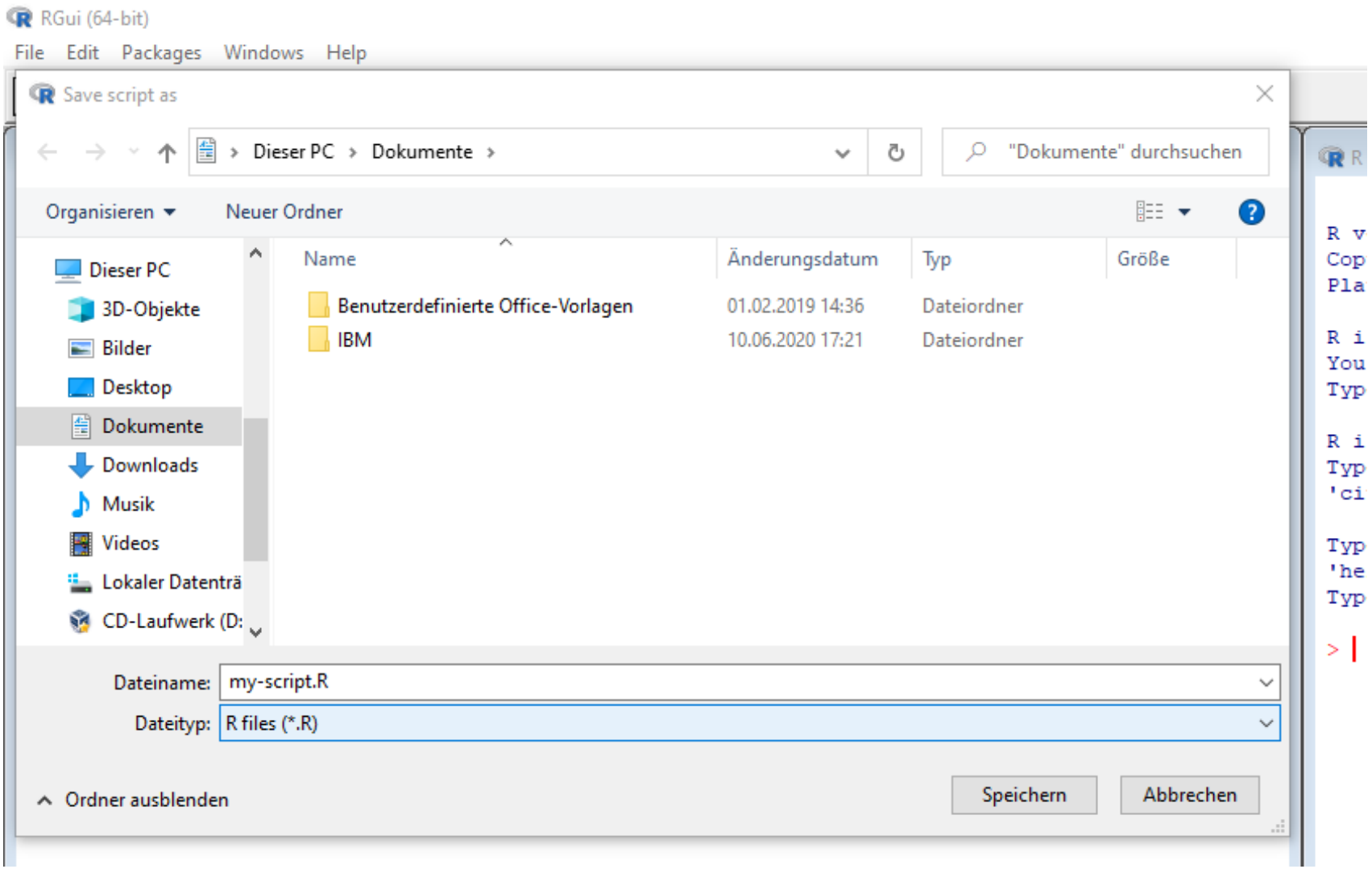
Recap: Using RGui



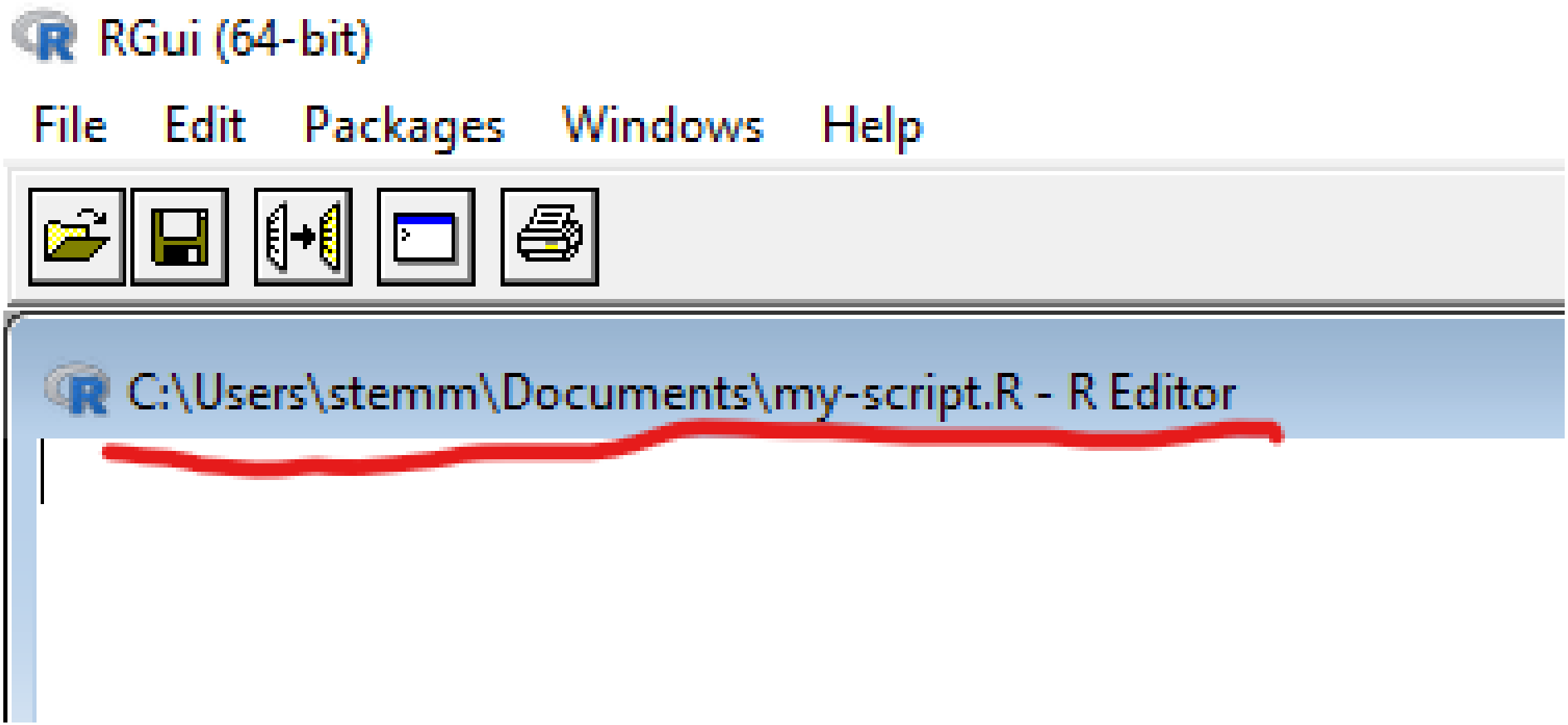
Recap: Using RGui



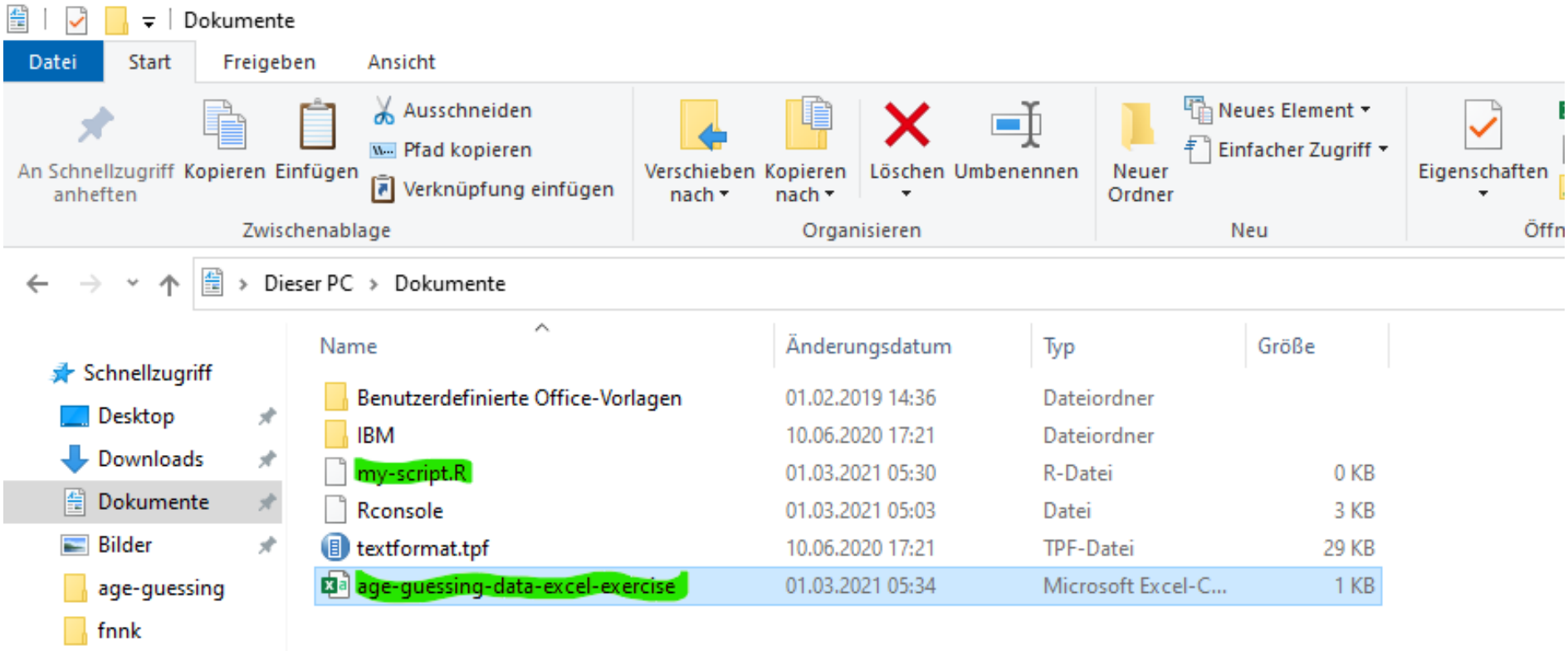
Recap: Using RGui



Recap: Using RGui



Recap: Using RGui

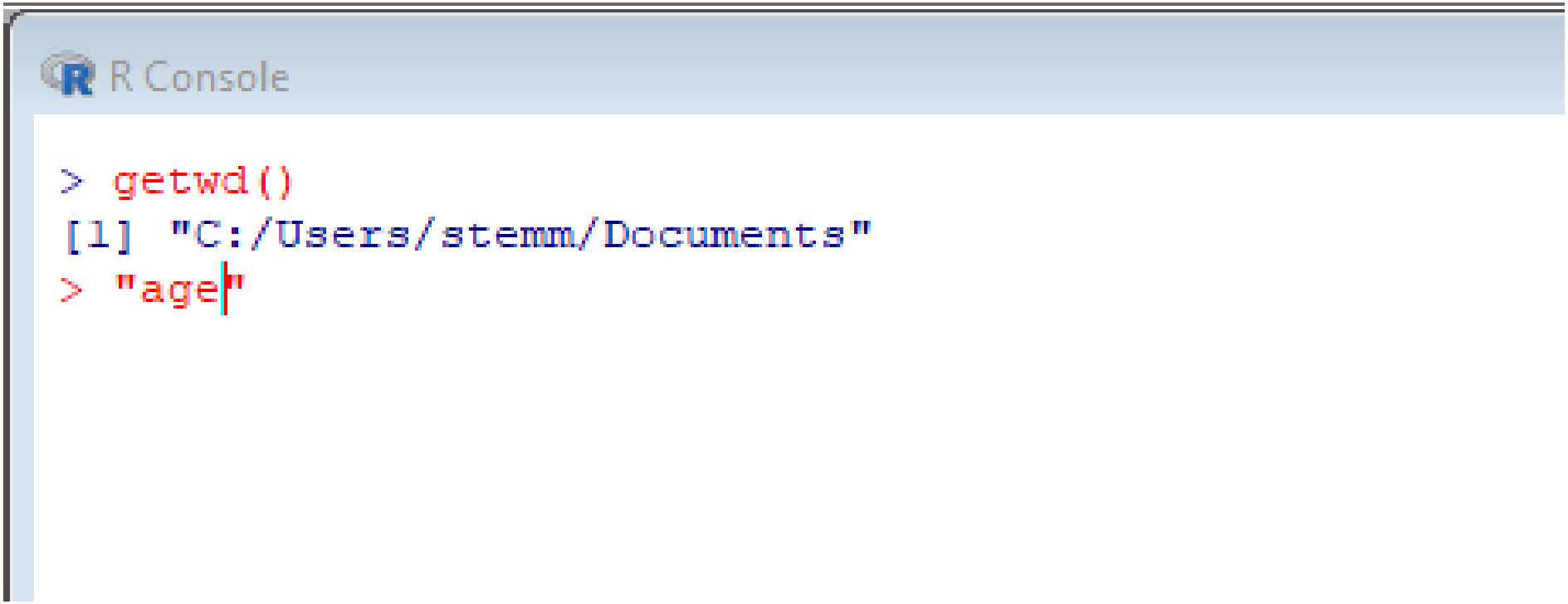


Recap: Using RGui

R Console

```
R version 4.0.3 (2020-10-10) -- "Bunny-Wunnies Freak Out"
Copyright (C) 2020 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> getwd()
[1] "C:/Users/stemm/Documents"
>
```

Recap: Using RGui



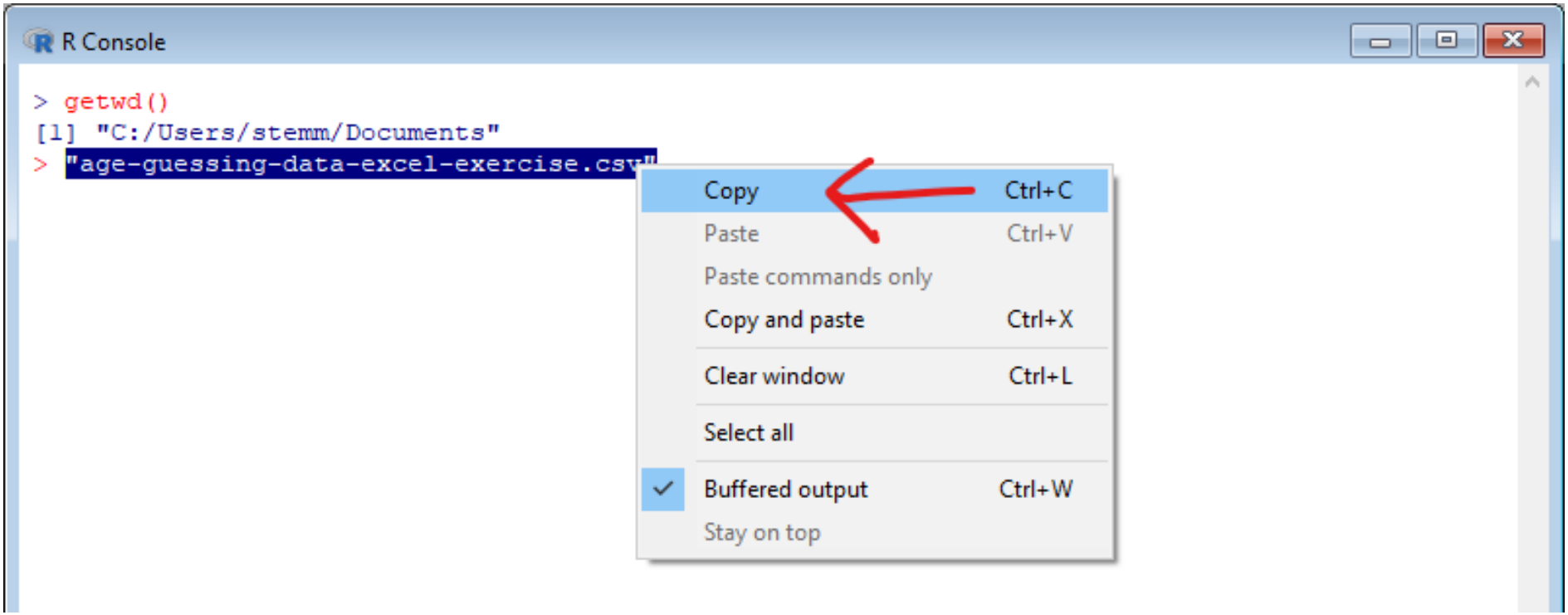
Press ⇤⇥ x2 to complete automatically

Recap: Using RGui

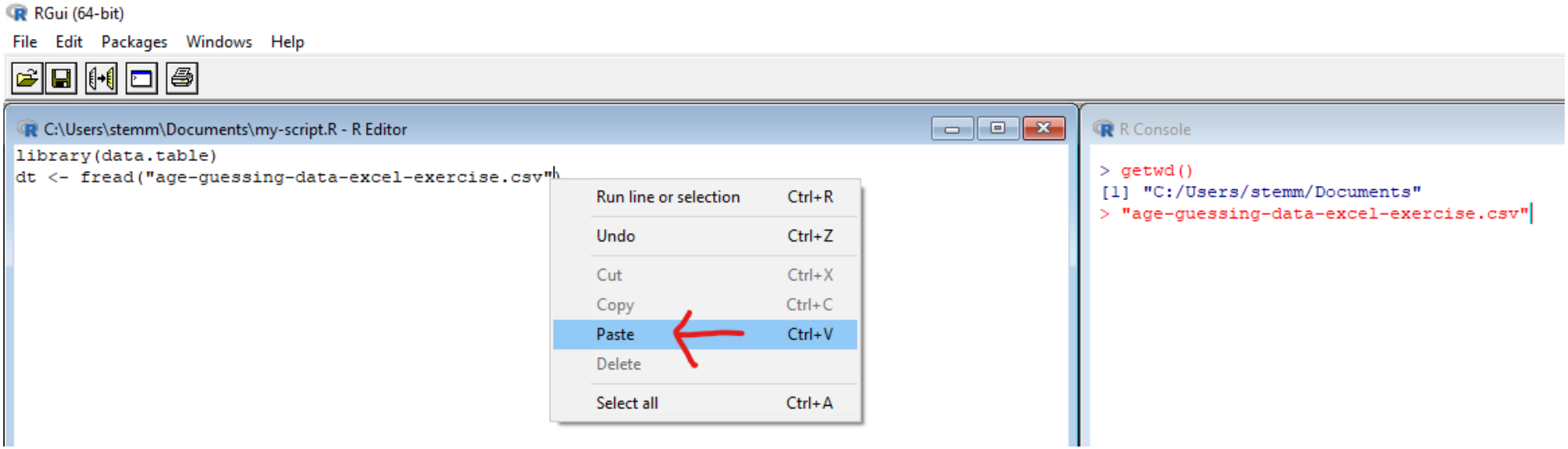
R Console

```
> getwd()
[1] "C:/Users/stemm/Documents"
> "age-guessing-data-excel-exercise.csv"
```

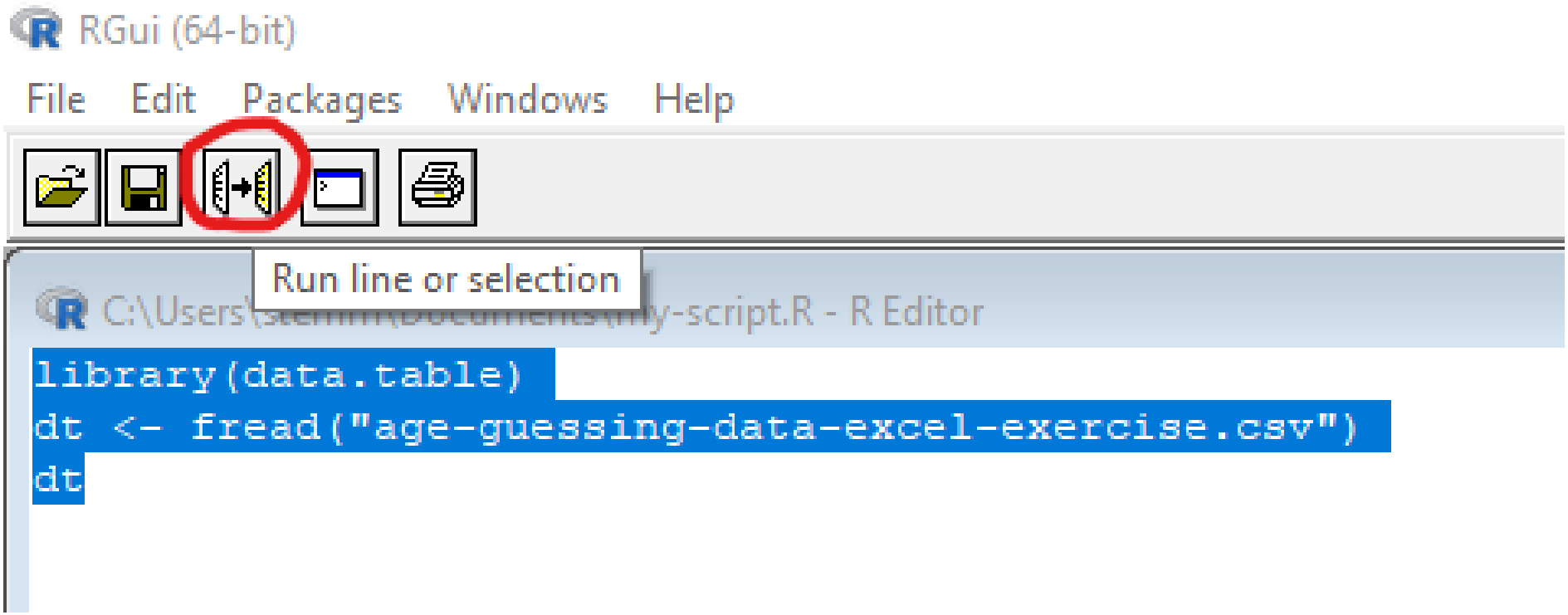
Recap: Using RGui



Recap: Using RGui



Recap: Using RGui



Recap: Using RGui

R Console

```
> library(data.table)
> dt <- fread("age-guessing-data-excel-exercise.csv")
> dt
    group card_1 card_2 card_3 card_4 card_5 card_6 card_7 card_8 card_9 card_10
 1:     A    -11    -15     11     -2      0      5     -4      1     -7      -6
 2:     B     -1    -12      4      7      7     -5     12      9      5       3
 3:     C      3    -12     20      9     23      1     12      7      5       3
 4:     D     -6    -10     -1     -3    -20      0    -11     -5     -3      -1
 5:     E      3    -12     10      8     13     -6      1     -3      6       4
 6:     F      0     13      1      0     20      7      0      1      1       2
 7:     G     -2     -6      0     -4    -10     -1     -8    -10     -5      -2
 8:     H      3      9      4      3      9      5      1      0      1       1
 9:     I     -8      2     -3      6     13     -1      6      9      5       3
10:     J     -4    -10      2     11     -4     -5    -11      8     -1       1
> 
```

Section 2

Recap: Working with data sets

Recap: Working with data sets

- Prepare data set: from Excel (`*.xlsx`) to CSV (`*.csv`)
- Read-in data set: `fread(file = ...)`
- Analyse, modify, plot data sets
- Save modified data set: `fwrite(x=..., file=..., sep=',')`

Section 3

Accessing variables from a data.table

Column Access - $

Using $ to access a column

```
library(data.table)
dt <- fread("age-guessing-data-excel-exercise.csv")
dt$group
```

```
##  [1] "A" "B" "C" "D" "E" "F" "G" "H" "I" "J"
```

```
dt$card_1
```

```
##  [1] -11  -1   3  -6   3   0  -2   3  -8  -4
```

R-console: Press ⇤⇥ behind $ to list all available column names

Selecting multiple columns

We can also use [] to access columns

```
dt[, c("group", "card_1")]
```

```
##     group card_1
##  1:     A    -11
##  2:     B     -1
##  3:     C      3
##  4:     D     -6
##  5:     E      3
##  6:     F      0
##  7:     G     -2
##  8:     H      3
##  9:     I     -8
## 10:     J     -4
```

Selecting multiple columns

. . . or a combination of both . . .

```
dt[, c("group", "card_1")]$group
```

```
##  [1] "A" "B" "C" "D" "E" "F" "G" "H" "I" "J"
```

Section 4

Accessing subsets from data.table

Accessing subsets from data.table

Subset: Selection of rows by row number

```
# first row
dt[1]
```

```
##    group card_1 card_2 card_3 card_4 card_5 card_6
## 1:     A    -11    -15     11     -2      0      5
##    card_7 card_8 card_9 card_10
## 1:     -4      1     -7      -6
```

Accessing subsets from data.table

Subset: Selection of rows by row number

```
# first three rows
dt[1:3]
```

```
##    group card_1 card_2 card_3 card_4 card_5 card_6
## 1:     A    -11    -15     11     -2      0      5
## 2:     B     -1    -12      4      7      7     -5
## 3:     C      3    -12     20      9     23      1
##    card_7 card_8 card_9 card_10
## 1:     -4      1     -7     -6
## 2:     12      9      5      3
## 3:     12      7      5      3
```

Accessing subsets from data.table

```
# all but the row 1 to 5
dt[-(1:5)]
```

```
##    group card_1 card_2 card_3 card_4 card_5 card_6
## 1:     F      0     13      1      0     20      7
## 2:     G     -2     -6      0     -4    -10     -1
## 3:     H      3      9      4      3      9      5
## 4:     I     -8      2     -3      6     13     -1
## 5:     J     -4    -10      2     11     -4     -5
##    card_7 card_8 card_9 card_10
## 1:      0      1      1       2
## 2:     -8    -10     -5      -2
## 3:      1      0      1       1
## 4:      6      9      5       3
## 5:    -11      8     -1       1
```

Accessing subsets from data.table

Subset: Selection of rows by variable value

```
# Only rows where group equals "A" or "D"
dt[group == "A" | group == "D"]
```

```
##    group card_1 card_2 card_3 card_4 card_5 card_6
## 1:     A    -11    -15     11     -2      0      5
## 2:     D     -6    -10     -1     -3    -20      0
##    card_7 card_8 card_9 card_10
## 1:     -4      1     -7      -6
## 2:    -11     -5     -3      -1
```

Symbol | is a logical "or": group can be "A" or "D"

Accessing subsets from data.table

Subset: Selection of rows by variable value

```
# Only rows where the absolute value
# of error is larger than 5
dt[abs(card_1) > 5]
```

```
##    group card_1 card_2 card_3 card_4 card_5 card_6
## 1:     A    -11    -15     11     -2      0      5
## 2:     D     -6    -10     -1     -3    -20      0
## 3:     I     -8      2     -3      6     13     -1
##    card_7 card_8 card_9 card_10
## 1:     -4      1     -7     -6
## 2:    -11     -5     -3     -1
## 3:      6      9      5      3
```

Accessing subsets from data.table

Subset: Selection of rows by variable value

```
# Only rows where group is one of "A", "B" or "C"
# and the absolute value of error for the 1st photograph
# is larger than 5
dt[group %in% c("A", "B", "C") & abs(card_1) > 5]
```

```
##    group card_1 card_2 card_3 card_4 card_5 card_6
## 1:     A    -11    -15     11     -2      0      5
##    card_7 card_8 card_9 card_10
## 1:     -4      1     -7      -6
```

- `group %in% c("A", "B", "C")` means `group == "A" | group == "B" | group == "C"`
- Symbol `&` is a logical "and", i.e. group must be "A", "B" or "C" AND error is larger than 5

Accessing subsets from data.table

Subset: Selection of rows by variable value

```
# same as above, but using logical OR instead of AND
dt[group %in% c("A", "B", "C") | abs(card_1) > 5]
```

```
##    group card_1 card_2 card_3 card_4 card_5 card_6
## 1:     A    -11    -15     11     -2      0      5
## 2:     B     -1    -12      4      7      7     -5
## 3:     C      3    -12     20      9     23      1
## 4:     D     -6    -10     -1     -3    -20      0
## 5:     I     -8      2     -3      6     13     -1
##    card_7 card_8 card_9 card_10
## 1:     -4      1     -7      -6
## 2:     12      9      5       3
## 3:     12      7      5       3
## 4:    -11     -5     -3      -1
## 5:      6      9      5       3
```

Section 5

Sorting

Sorting

To sort means to change the order of rows

Sorting: ascending vs descending

```
# ascending
dt[order(card_2)]
```

```
##      group card_1 card_2 card_3 card_4 card_5 card_6
##  1:      A    -11    -15     11     -2      0      5
##  2:      B     -1    -12      4      7      7     -5
##  3:      C      3    -12     20      9     23      1
##  4:      E      3    -12     10      8     13     -6
##  5:      D     -6    -10     -1     -3    -20      0
##  6:      J     -4    -10      2     11     -4     -5
##  7:      G     -2     -6      0     -4    -10     -1
##  8:      I     -8      2     -3      6     13     -1
##  9:      H      3      9      4      3      9      5
## 10:      F      0     13      1      0     20      7
##      card_7 card_8 card_9 card_10
```

Sorting

To sort means to change the order of rows

Sorting: ascending vs descending

```
# descending
dt[order(card_2, decreasing = TRUE)]
```

```
##      group card_1 card_2 card_3 card_4 card_5 card_6
##  1:      F      0     13      1      0     20      7
##  2:      H      3      9      4      3      9      5
##  3:      I     -8      2     -3      6     13     -1
##  4:      G     -2     -6      0     -4    -10     -1
##  5:      D     -6    -10     -1     -3    -20      0
##  6:      J     -4    -10      2     11     -4     -5
##  7:      B     -1    -12      4      7      7     -5
##  8:      C      3    -12     20      9     23      1
##  9:      E      3    -12     10      8     13     -6
## 10:      A    -11    -15     11     -2      0      5
##      card_7 card_8 card_9 card_10
```

Section 6

Summary: What functions did we learn?

Summary: What functions did we learn?

- `dt$column`: get column from data.table
- `dt[, c("columnA", "columnB")]`: get one or more columns from data.table
- `order(column)`: Sort values in `column` in ascending order
- `order(column, decreasing = TRUE)`: Sort values in `column` in decreasing order
- `A %in% c(B, C; D)`: shortcut for A = B or A = C or A = D
- Logical Or-operator |: A=="normal" | A=="small"
- Logical And-operator &: `A>1 & A<5`

Section 7

Exercises

Age Guessing

Analyse the age guessing data set. For each question, write one line of R code to determine the answer.

1. Which team had the lowest absolute error for card no. 8?
2. Which photograph card was the most difficult one, i.e. the one with the highest total error

Bird Flue - Avian Influenza A/ H5N1

Zuur et al. (2009): The file `BirdFlu.xls` contains the annual number of confirmed cases of human Avian Influenza A/(H5N1) for several countries reported to the World Health Organization (WHO). The data were taken from the WHO website (www.who.int/en/).

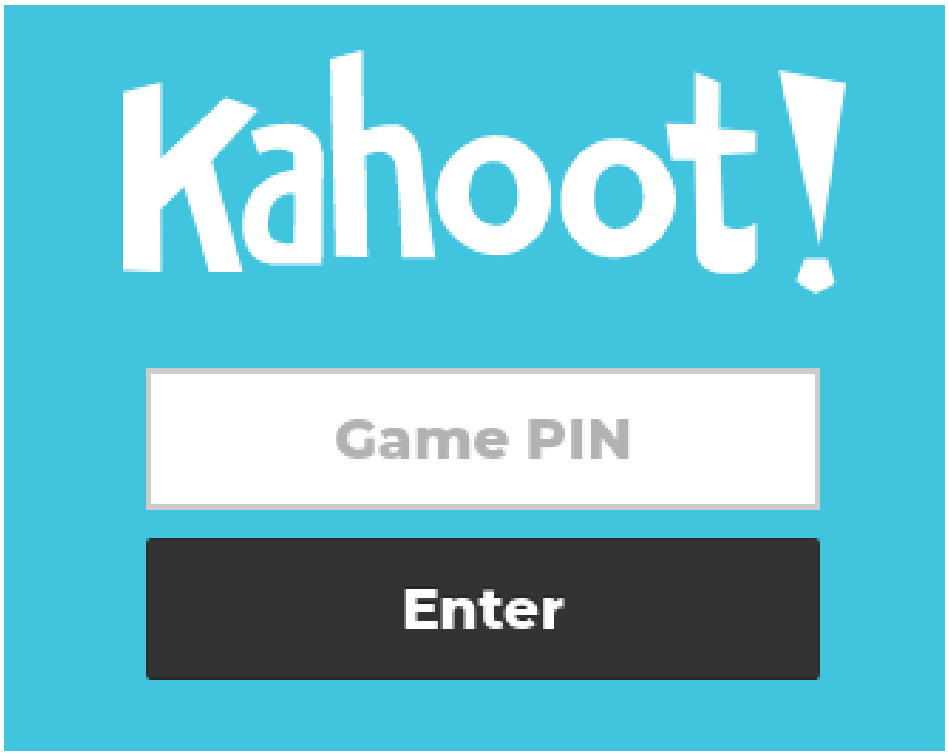
Download the data set and copy file `BirdFlu.xls` from http://highstat.com/Books/Book3/MoreData.zip. Convert `BirdFlu.xls` into a csv format, ie. `birdflu.csv`. Write a R-script to answer the following questions:

1. What is the total number of bird flu cases in 2003 and in 2005?
2. Which country has had the *most* cases?
3. Which country has had the *least* bird flu deaths?

Section 8

Quiz

Quiz



- Phone or Computer: https://www.kahoot.it
- Wifi: Platinum
- Password: H2SvsH2O

Alain Zuur, Elena N Ieno, and Erik Meesters. *A Beginner's Guide to R*. Springer Science & Business Media, 2009.