



# DATA-DRIVEN DECISION MAKING IN EDUCATION

*An introduction to R on PISA datasets*



Agnes Salanki

June 2019



Hotels.com™

## Before we start

**1. Presentation + dataset:**  
[http://bit.ly/wosr\\_slides](http://bit.ly/wosr_slides)

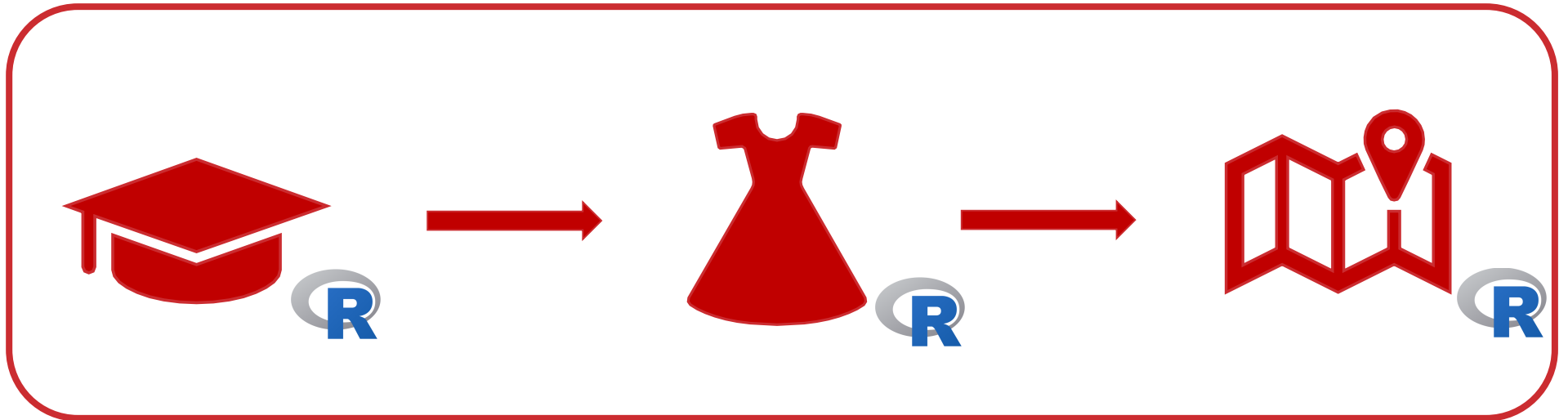
**2. Go to**

<https://rstudio.cloud>

**and create an account**



# Who am I?



Presentation + dataset: [http://bit.ly/wosr\\_slides](http://bit.ly/wosr_slides)

Go to <https://rstudio.cloud> and create an account



# What is R?

- Programming language/software environment
- Statistics and visualization



Presentation + dataset: [http://bit.ly/wosr\\_slides](http://bit.ly/wosr_slides)

Go to <https://rstudio.cloud> and create an account



# Why am I talking about R here?

**I love**

- **the language**
- **the community**
- **sharing the word**



Presentation + dataset: [http://bit.ly/wosr\\_slides](http://bit.ly/wosr_slides)

Go to <https://rstudio.cloud> and create an account



# Outline of the workshop





# Packages

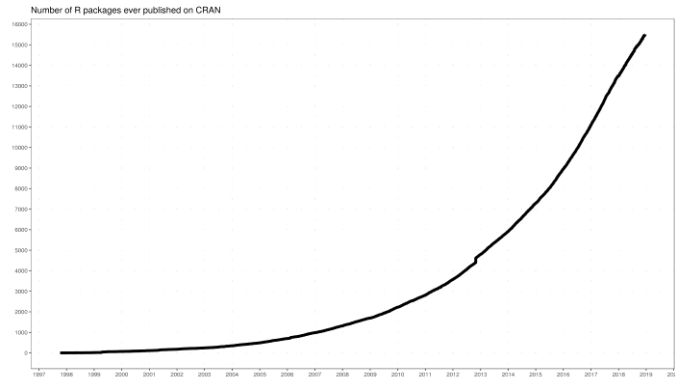


Hotels.com™ 7

# Why is it so easy to work with R?

## CRAN

Comprehensive R Archive Network



```
library(numbers)
primeFactors()
```

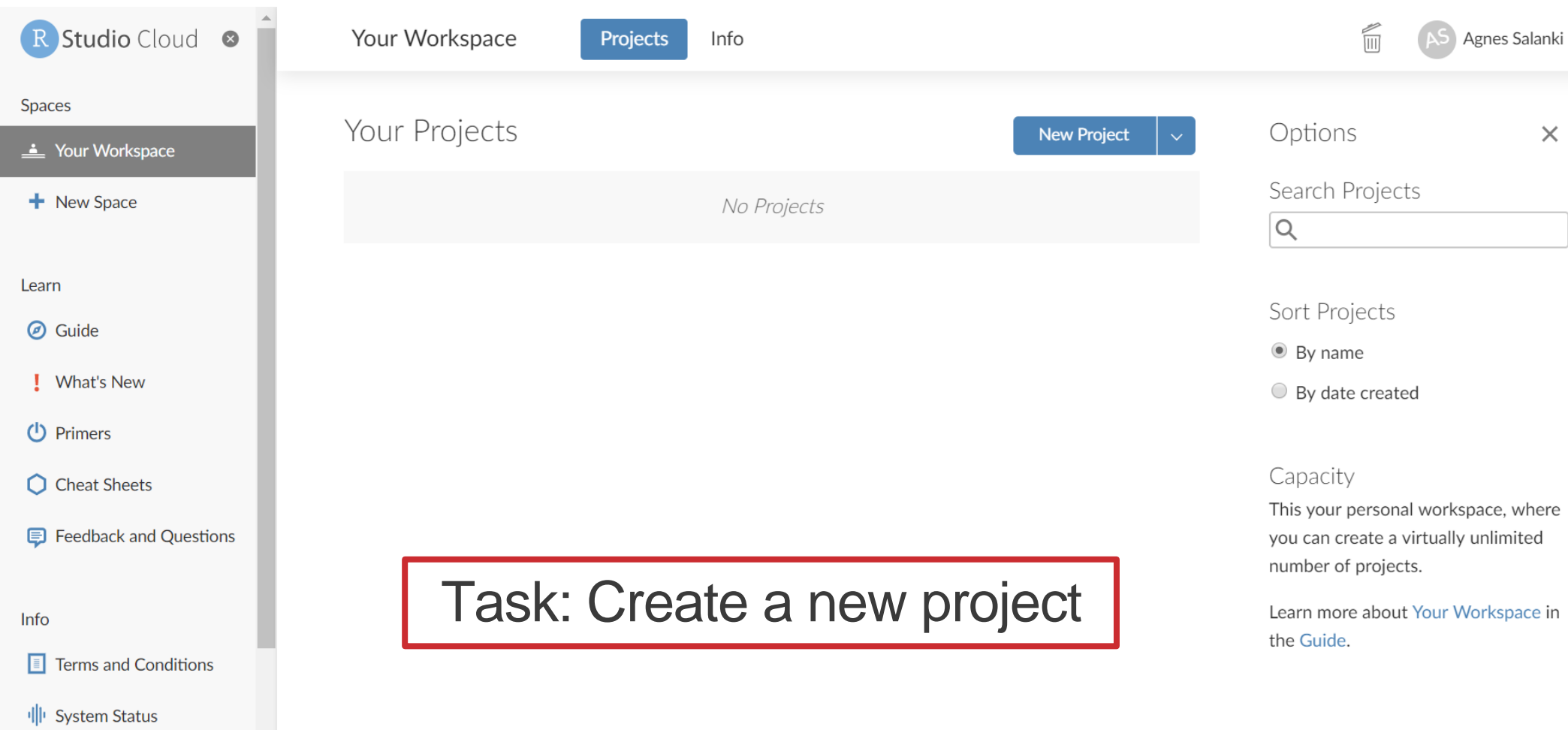
```
library(fun)
mine_sweeper()
```

●	2	1	2	2	2	2	1	1	0
1	2	●	2	●	●	3	●	2	0
0	2	2	3	2	2	4	●	4	1
0	1	●	2	1	0	2	●	4	●
0	2	3	●	2	1	1	1	3	●
1	2	●	3	●	1	0	1	3	3
●	2	1	2	1	1	1	2	●	●
1	1	0	1	2	2	2	●	4	3
0	0	0	1	●	●	2	2	●	1
0	0	0	1	2	2	1	1	1	1

Go to <https://rstudio.cloud> and create an account



# Why is it so easy to work with R?



The screenshot displays the R Studio Cloud interface. On the left is a sidebar with the following sections:

- Spaces:** Includes 'Your Workspace' (selected) and a '+ New Space' button.
- Learn:** Includes links for 'Guide', 'What's New', 'Primers', 'Cheat Sheets', and 'Feedback and Questions'.
- Info:** Includes links for 'Terms and Conditions' and 'System Status'.

The main content area is titled 'Your Projects' and features a 'New Project' button. Below the title, it states 'No Projects'.

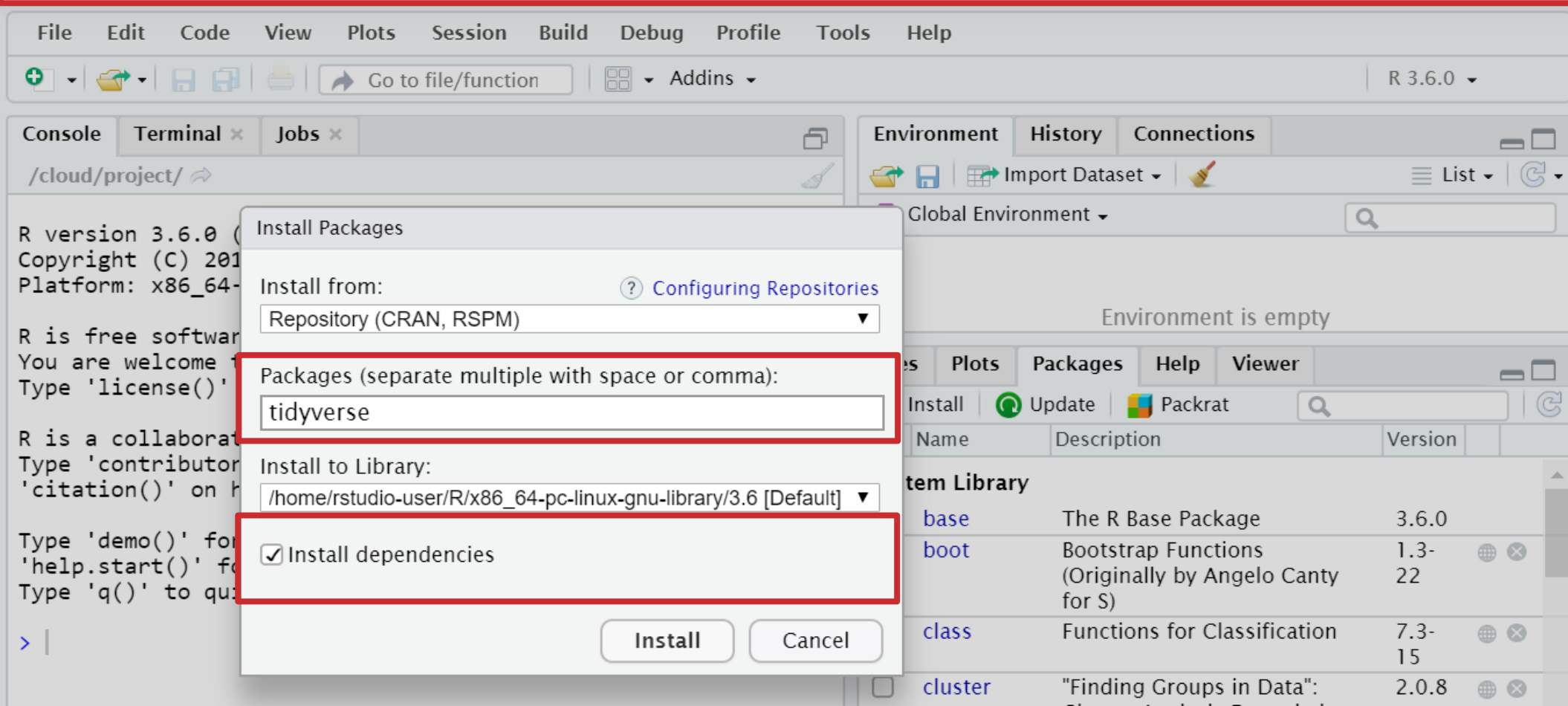
On the right side, there is a panel with the following options:

- Options:** A close button (X).
- Search Projects:** A search input field with a magnifying glass icon.
- Sort Projects:** Two radio button options: 'By name' (selected) and 'By date created'.
- Capacity:** A section explaining that this is the personal workspace for creating a virtually unlimited number of projects, with a link to the 'Guide' for more information.

At the bottom center, a red-bordered box contains the text: **Task: Create a new project**

# Why is it so easy to work with R?

## Task: Install the tidyverse package group



The screenshot shows the RStudio interface with the 'Install Packages' dialog box open. The dialog box has the following fields and options:

- Install from:** Repository (CRAN, RSPM) (with a link to 'Configuring Repositories')
- Packages (separate multiple with space or comma):** tidyverse
- Install to Library:** /home/rstudio-user/R/x86\_64-pc-linux-gnu-library/3.6 [Default]
- ☒ Install dependencies

The background shows the RStudio console with the R version 3.6.0 and the Environment pane showing the Global Environment is empty. The Packages pane shows the installed packages in the system library.

Name	Description	Version
base	The R Base Package	3.6.0
boot	Bootstrap Functions (Originally by Angelo Canty for S)	1.3-22
class	Functions for Classification	7.3-15
cluster	"Finding Groups in Data":	2.0.8



# Community



# Where can I get help?

Task: load the tidyverse package group RUNNING  
*library(tidyverse) (Ctrl + Enter)*

## Meetups

e.g., LondonR,  
R-Ladies London

## Conferences

e.g., satRdays

## Help pages

?tidyverse  
??tidyverse

## Stackoverflow

+ Github  
+blogs

## Twitter

#rstats

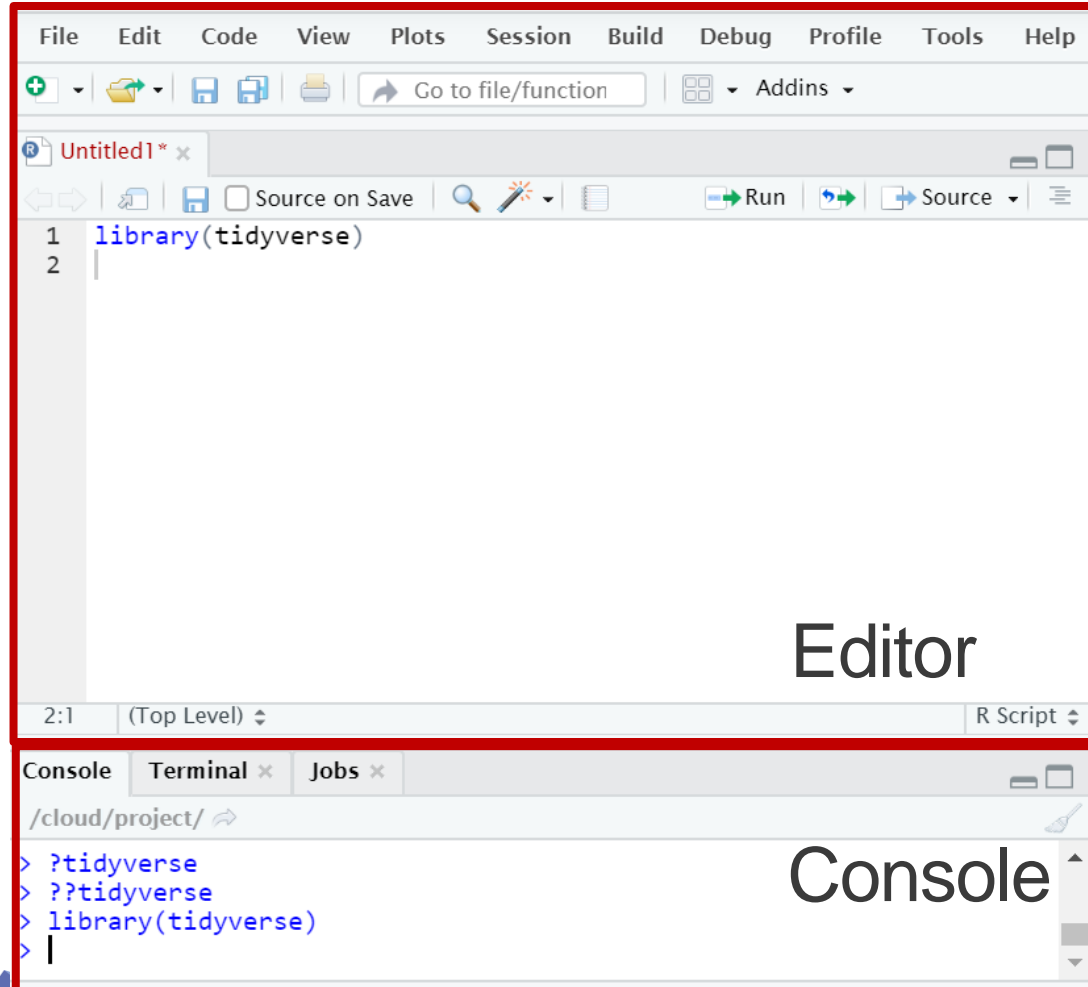




# RStudio



Hotels.com™ 13



The screenshot shows the RStudio Editor window. The top menu bar includes File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, and Help. Below the menu is a toolbar with icons for file operations and a search bar labeled 'Go to file/function'. The editor pane shows a file named 'Untitled1\*' with two lines of R code: `1 library(tidyverse)` and `2`. The status bar at the bottom indicates the cursor is at line 2, column 1, and the file is an R Script.

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function

Untitled1\* x

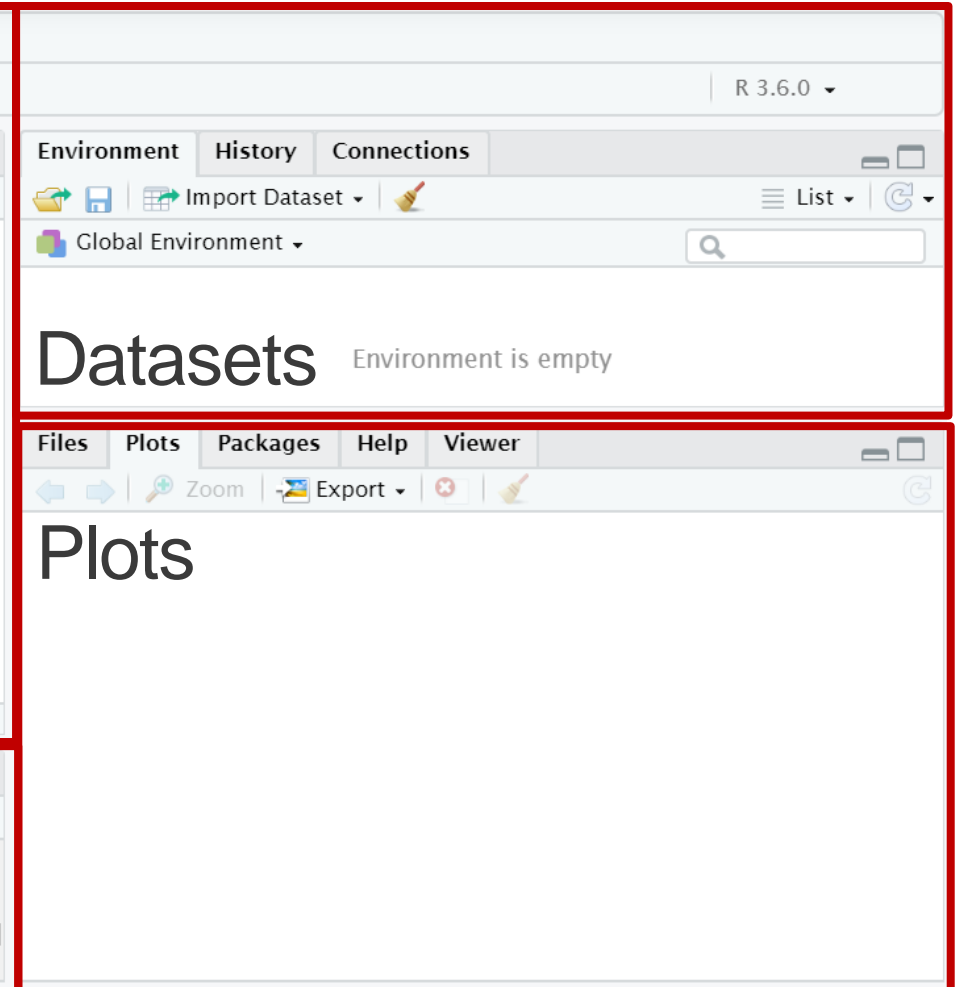
```
1 library(tidyverse)
2
```

Source on Save Run Source

2:1 (Top Level) R Script

Editor

Console



The screenshot shows the RStudio Environment and Plots panels. The top panel is the Environment pane, which shows the Global Environment and a search bar. The bottom panel is the Plots pane, which is currently empty. The status bar at the bottom indicates the R version is 3.6.0.

R 3.6.0

Environment History Connections

Import Dataset

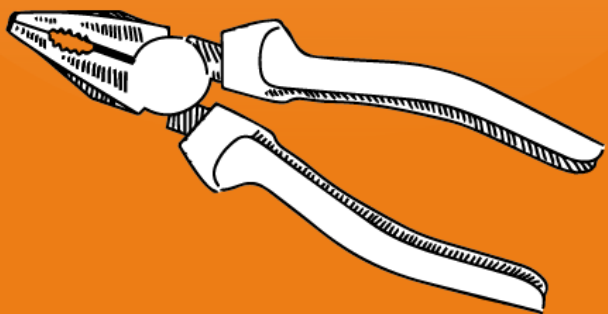
Global Environment

Datasets Environment is empty

Files Plots Packages Help Viewer

Zoom Export

Plots



dplyr

[www.rstudio.com](http://www.rstudio.com)

# Data Transformation

dplyr



# Data description: a data frame

**Task: Load the dataset into memory**

```
library(tidyverse)
```

```
pisa <- read_csv("http://bit.ly/wosr_pisa_data")
```

```
head(pisa)
```

```
## or
```

```
view(pisa)
```

```
## and
```

```
summary(pisa)
```





# Context: PISA tests

- Programme for International Student Assessment
- 15-year-old students
- Competence in three fields: maths, science and reading
- Questionnaire about the students, their schools, parents, household, etc.
- Results and data are published in the following year
- UK 2015 findings: <http://www.oecd.org/pisa/PISA-2015-United-Kingdom.pdf>

# Data transformation primitives

**Task: How many children got tested in each country?**

**SELECT** columns

**FROM** table

**WHERE** condition

**GROUP BY** columns

**ORDER BY** columns

```
pisa <- read_csv("http://bit.ly/wosr_pisa_data")
```

# Data transformation primitives

**Task: How many children got tested in each OECD country?**

select

**SELECT** CNT, COUNT(1) as records

**FROM** pisa

%>%

summarise

filter

**WHERE** OECD = 'Yes'

**GROUP BY** CNT

group\_by

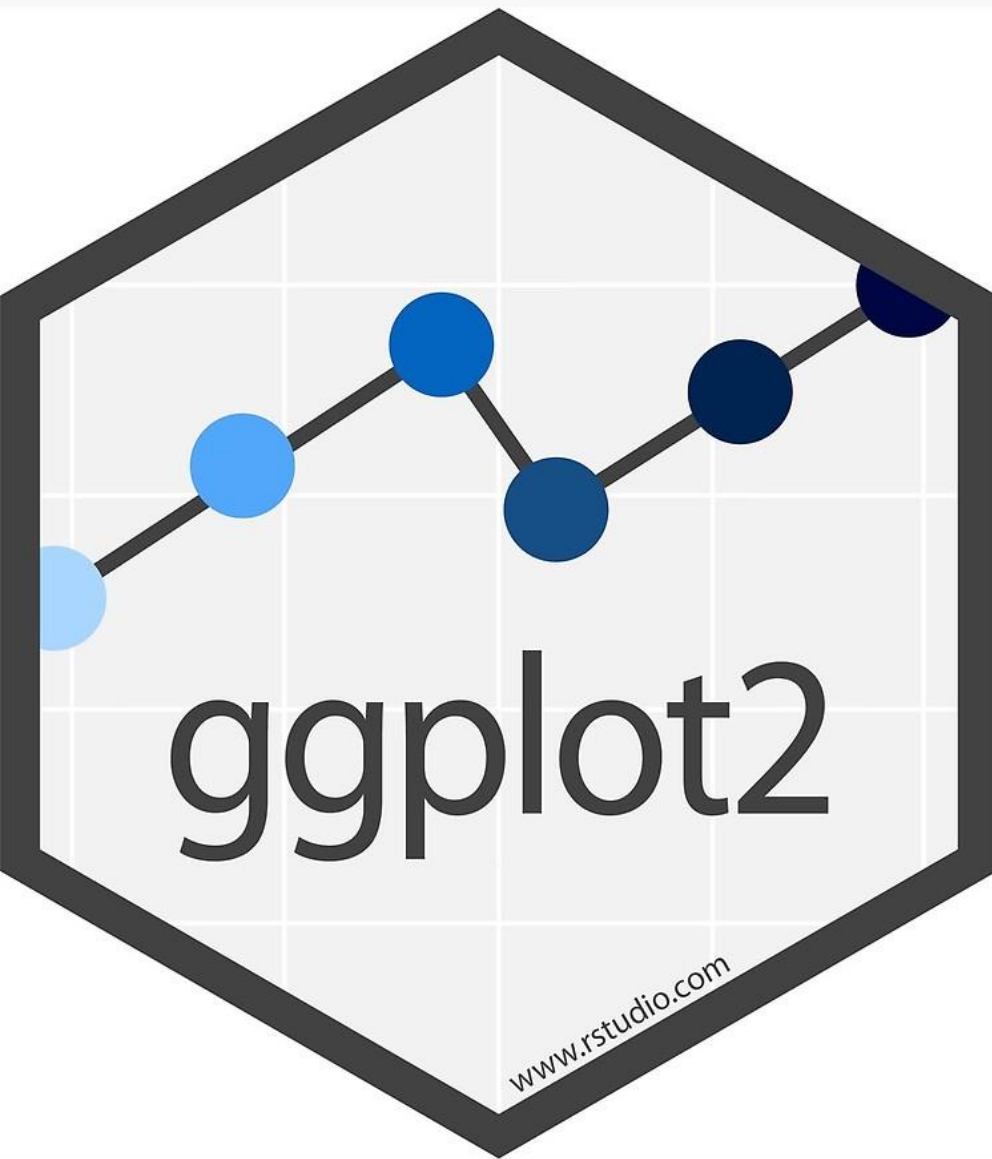
arrange

**ORDER BY** records

# Data transformation primitives

Statement: *“By age 15, students in the United Kingdom perform above the OECD average in science (509 score points) and reading (498 points) and around the OECD average in mathematics (492 points).”*

**Task: Calculate the median science result in each country**



# Data Visualization

ggplot2



# Approach of layers

Choose a data frame



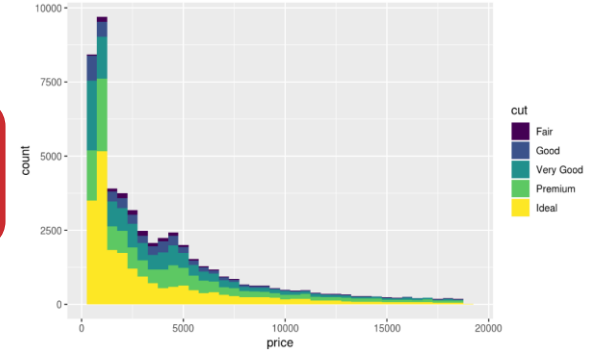
Choose a chart type



Choose the column to aesthetics mapping



Set the visual parameters



# Approach of layers

Choose a data frame

`ggplot()`

+

Choose a chart type

`geom_*()`

+

Choose the column to aesthetics mapping

`aes(x = ...,  
y = ...)`

+

Set the visual parameters

`theme(),  
scale_*()`



# Approach of layers

Statement: *“By age 15, students in the United Kingdom perform above the OECD average in science (509 score points) and reading (498 points) and around the OECD average in mathematics (492 points).”*

**Task: Plot the median science score for each OECD country**



# Approach of layers

Statement: *“By age 15, students in the United Kingdom perform above the OECD average in science (509 score points) and reading (498 points) and around the OECD average in mathematics (492 points).”*

**Task: Using boxplots, visualize the distribution of science scores in each OECD country**

# Approach of layers

Statement: *“Students with an immigrant background (first or second generation) in the United Kingdom, as in many other OECD countries, do not perform as well in science as students without an immigrant background.”*

**Task: Using boxplots, visualize the distribution of science points in the UK for each immigrant status**

# Approach of layers

Statement: *“In the United Kingdom, boys and girls are equally likely to score at Level 5 or 6, the highest levels of proficiency, in science (12% of boys and 10% of girls) (Table I.2.6a).”*

**Task: Using boxplots, visualize the distribution of science points in the UK for each gender**

## Approach of layers

Statement: *“Even though gender differences in science performance tend to be small on average, in 33 countries and economies, the share of top performers in science is larger among boys than among girls. In the United Kingdom, as a whole, there is no significant difference in the share of top performers among boys and girls (Table I.2.6a), and this is also true in England, Northern Ireland, Scotland and Wales (Table B2.I.3).”*

**Task: Using boxplots, visualize the distribution of science points in the UK for each gender, separated for each region**

## Approach of layers

Statement: *“Even though gender differences in science performance tend to be small on average, in 33 countries and economies, the share of top performers in science is larger among boys than among girls. In the United Kingdom, as a whole, there is no significant difference in the share of top performers among boys and girls (Table I.2.6a), and this is also true in England, Northern Ireland, Scotland and Wales (Table B2.I.3).”*

**Task: Using boxplots, visualize the distribution of science points in OECD countries, using different colors for each gender**

# Approach of layers

**Task: Modify the previous plot to have**

- *white background*
- *meaningful axis labels,*
- *bold fonts and*
- *different colors*



# Applications



# A few companies using R





# Summary

