Data Management in R/RStudio

Previous Class

- Detailed description of R & RStudio installation process
- RStudio interface
  - Script window (top left)
  - Console window (bottom left)
  - Environment window (top right)
  - Miscellaneous window (bottom right)
- Creating a project in RStudio
- Creating scripts
- Writing and executing codes in RStudio
- Basic operations, data types, vectors, logical operations, loops, etc.

Announcements

- Students form groups of three on canvas (if you have not yet done) – Due date and time is today 8 pm
- Unassigned students are randomly grouped or allocated after 8 pm
- Homework 1 will be available from today (by 11:59 pm)
- **Due date & time** in the syllabus file
- Start exploring the Homework 1 problems

Today's agenda

- Exposure to different Data Science packages in R
- Basic Data Management
    - Reading data
    - Exploring data
    - Summarizing data
    - Other operations

Packages

- Fundamental unit of share-able code
- Bundles together code, data, documentation, and tests and provides an easy method to share with others
- 19,046 packages available on CRAN as of today
- Numerous packages catered to wide applications

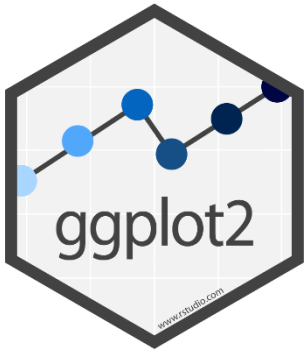
Tidyverse

- Collection of packages for Data Science in R

Data Import

Data Manipulation

Data Tidying



Graphics

Advanced Functions

Data frames

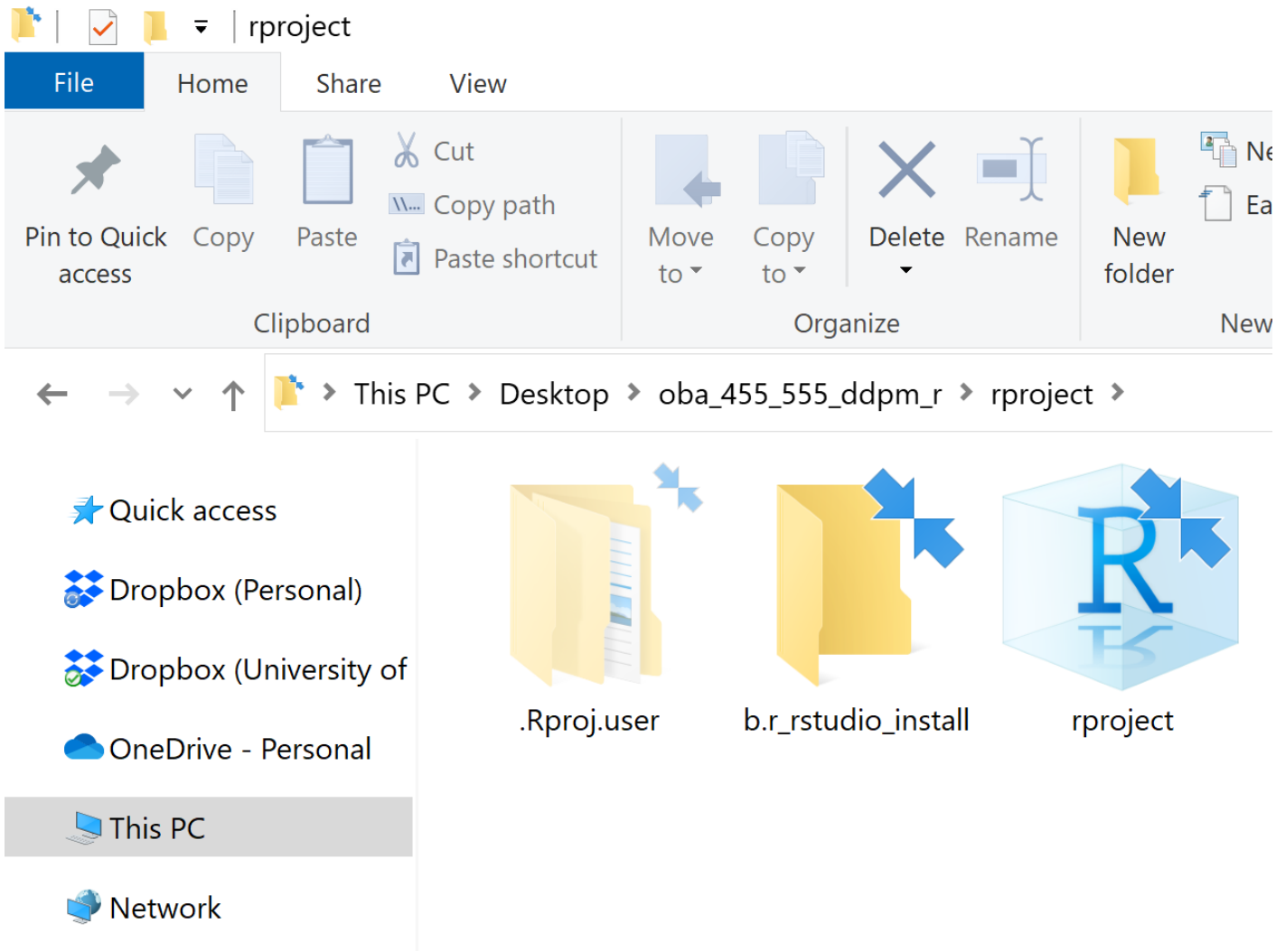
String manipulation

Categorical variables

- Explore https://www.tidyverse.org/ for more details

Open RStudio

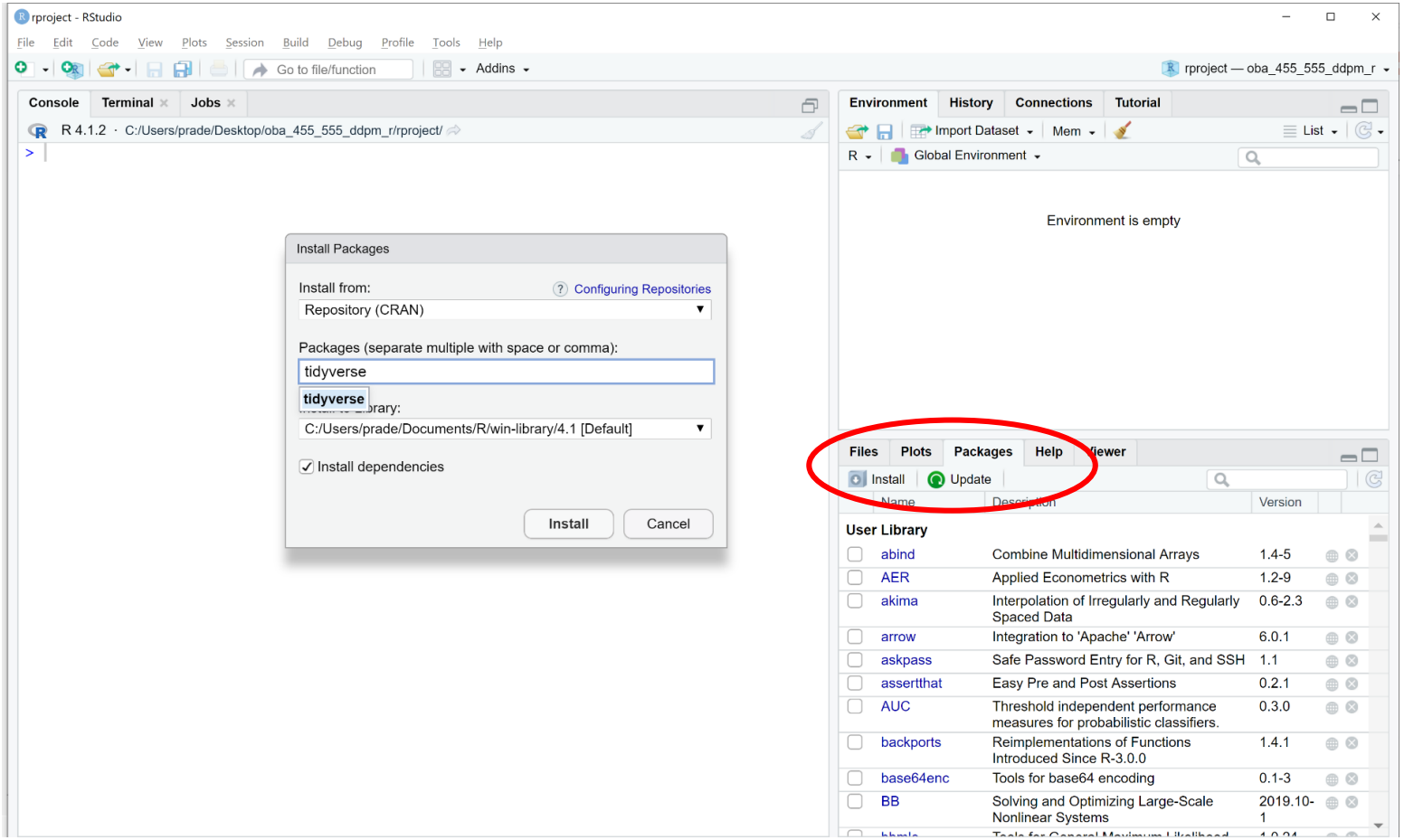
- Double click "**rproject**" icon



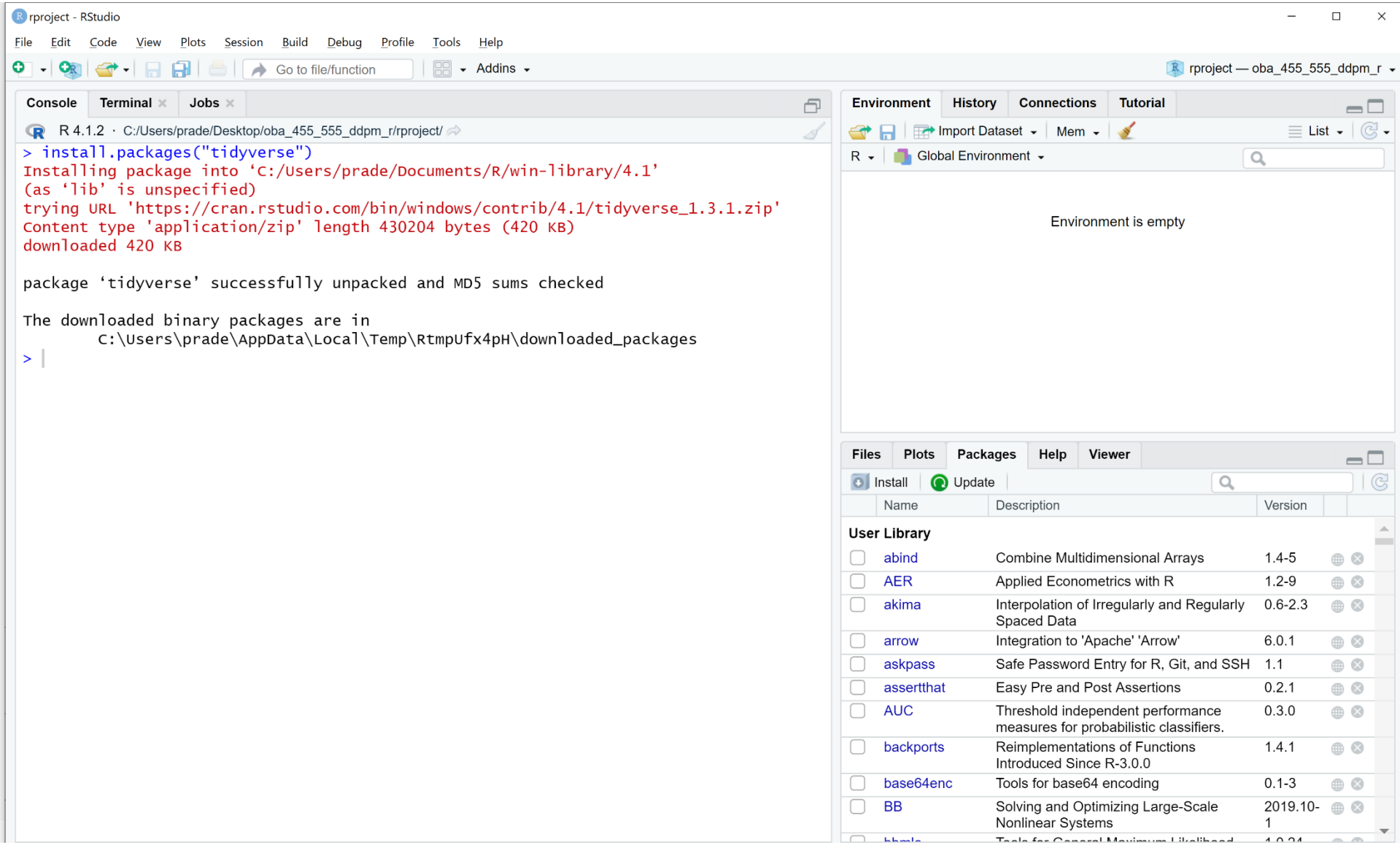
Installing Tidyverse packages in R

Installing Tidyverse packages

- Click "**Install**" icon in tab "**Packages**" on **Miscellaneous** window
- Type "**tidyverse**" under "**Packages"** and activate "**Install dependencies**"
- Click **"Install"**

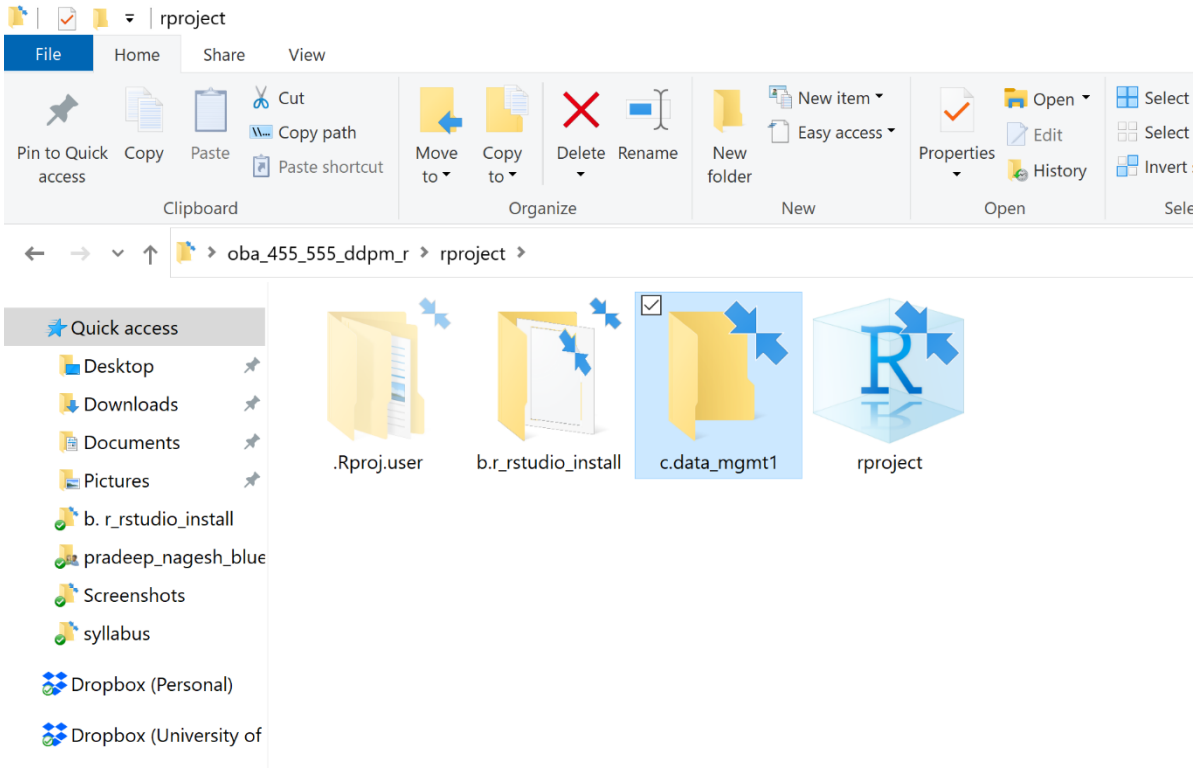


Installing Tidyverse packages



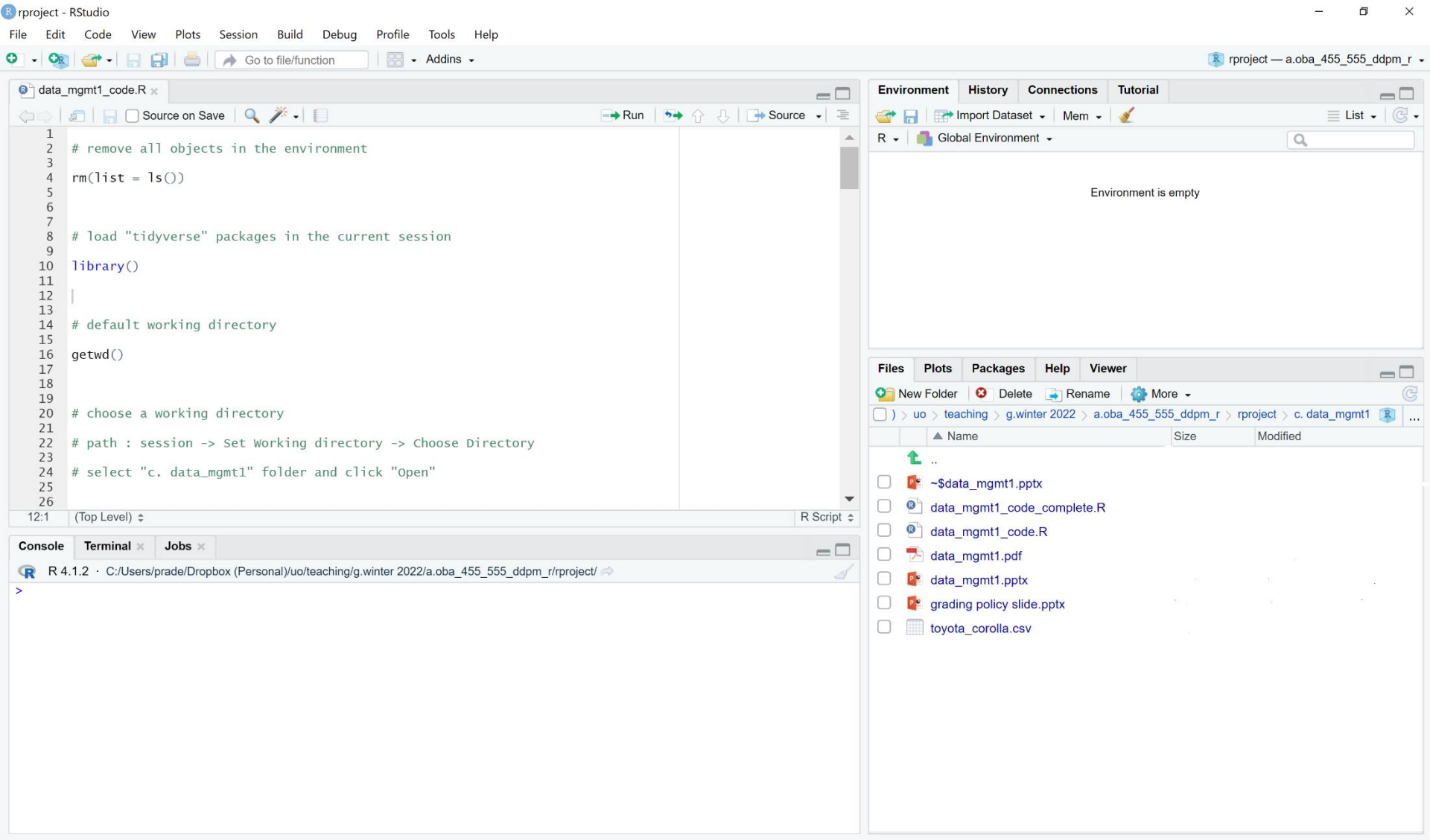
Mandatory Steps

- Create folder "**c. data_mgmt1**" within "**oba_455_555_ddpm_r/rproject**"



- Download "**data_mgmt1_code.R**" & "**toyota_corolla.csv**" files from canvas
- Place them in the path "**oba_455_555_ddpm_r/rproject/c.data_mgmt1**"
- Open "**data_mgmt1_code.R**" file within RStudio

Interface Display



Working directory

- When you read data, R looks into the working directory
- The default working directory location is known by code "**getwd( )**"
- How to define working directory or location where the data is present?

Session → Set Working Directory →

Choose Directory

- Select "**c.data_mgmt1**" folder and click "**Open**"

Data Import

Data Import

- Fast and friendly way to read rectangular data (csv, tsv, fwf ……)
- Most of the data comes in csv files
- Occasionally data comes in form of databases or other software format (SAS, STATA……..)

Commonly used functions

- **read_csv( )** : comma separated (CSV) files
- **read_tsv( )** : tab separated files
- **read_delim( )**: general delimited files
- **read_fwf( )** : fixed width files
- **read_table( )** : tabular files where columns are separated by white-space
- **read_log( )** : web log files
- Explore https://readr.tidyverse.org/ for more details

Data on used Toyota Corolla cars

- Data on sales of used cars in the Netherlands, late summer 2004
  - Each **column** is a **variable**
  - Each **row** is an **observation (or case)**
- Attributes
  - model : Toyota car model
  - Price : offer price in euros
  - age_08_04 : age in months as of august 2004
  - mfg_month : manufacturing month (1,2,3……12)
  - mfg_year : manufacturing year
  - km : accumulated kilometers on the odometer
  - fuel_type : fuel type (petrol, diesel, cng)
  - hp : horse power

Data on used Toyota Corolla cars cont.…

- Attributes
  - met_color : metallic colour (yes = 1, no = 0)
  - Color : colour, Automatic : automatic (yes = 1, no = 0)
  - cc : cylinder volume in cubic meters
  - doors, cylinders, gears, quarterly_tax
  - weight = weight in kilograms
  - mfr_guarantee : manufacturer guarantee
  - bovag_guarantee, guarantee_period
  - Abs, airbag_1, airbag_2, airco, automatic_airco,
  - Boardcomputer, cd_player, central_lock, powered_windows, power_steering, radio, mistlamps
  - sport_model, backseat_divider, metallic_rim, radio_cassette, parking_assistant, tow_bar

Commonly used functions

- **class(toyota)** : class of the object
- **str(toyota)** : structure of the object
- **View(toyota)** : opens a window with the data
- **head(toyota)** : displays first 6 rows (default) of data
- **nrow(toyota)** : produces an output of number of rows in the data
- **ncol(toyota)** : produces an output of number of columns in the data
- **dim(toyota)** : produces a vector of rows, columns in the data
- **summary(toyota)** : produce summary of all the variables in the data

Packages for reading other types of Data

- SPSS, Stata and SAS – **haven**
- Excel files (.xls and .xlsx) – **readxl**
- Databases – **DBI**
- JSON – **jsonlite**
- XML – **xml2**
- Web APIs – **httr**
- HTML (Web Scraping) – **rvest**

Data Manipulation

Data Manipulation

- dplyr is grammar of data manipulation
- Functions in this package help you solve the most common data manipulation challenges

Commonly used functions

- **mutate( )** : adds new variables that are functions of existing variables or replaces the existing variables with values of your choice
- **select( )** : picks variables based on their names
- **filter( )** : picks observations based on their values
- **summarise( )** : reduces multiple values down to single summary
- **arrange( )** : change the ordering of observations
- Explore https://dplyr.tidyverse.org/ for more details

Pipe operator

- Symbol "**%>%**" in R is called the pipe operator
- Pipe operators are used to perform stated actions on the data
- E.g. **toyota %>% summarise(avg_price = mean(price))**
- The casual meaning of "**%>%**" is "**do the action**"

Data Manipulations

- Summarize observations
  - Generate summary statistics - mean, median, quantile, variance, standard deviation
- Summarize group observations
  - Generate summary statistics by a group (e.g., gender, age group……)
- Manipulate observations
  - Filtering, sampling selected observations
- Order observations
  - Sorting the data by ascending or descending order of a variable
- Manipulate variables
  - Creating new variables or columns

Next Class

- Advanced Data Management & Graphics in R/RStudio
- Advanced Operations
  - Tidying
  - Binding
  - Appending
  - Merging
  - Long ↔ Wide
  - ……
- Graphics
  - Histogram, Bar chart
  - Scatter plot, Boxplot

Thank You