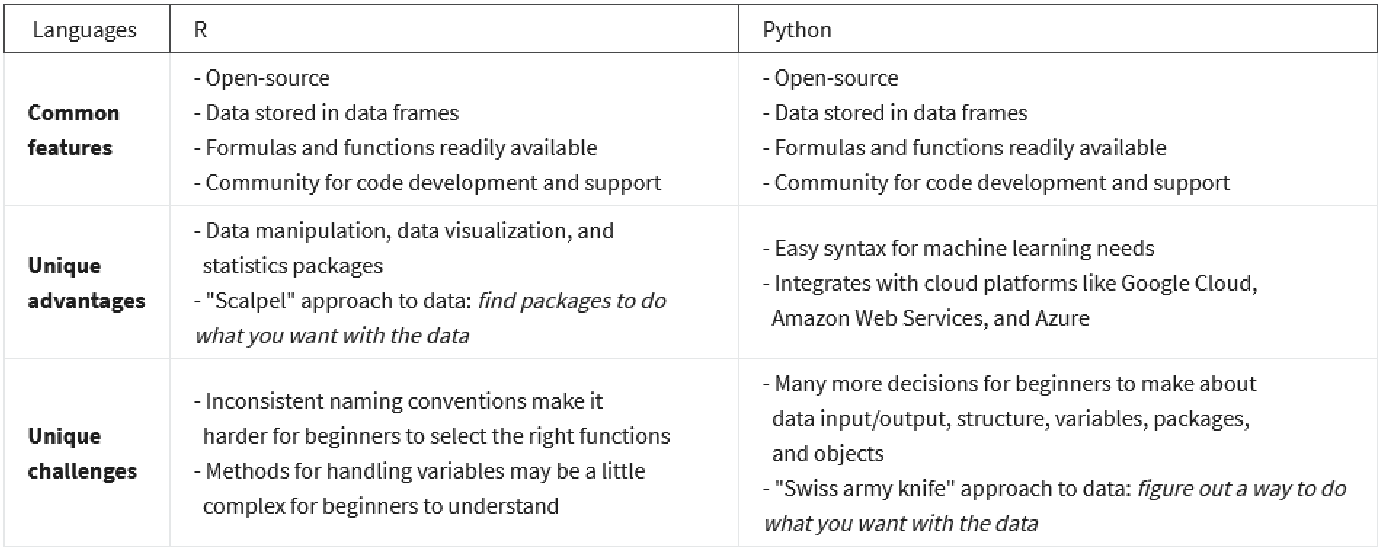
Google data analytics professional course

Week -1

The exciting world of programming

The R-versus-Python debate



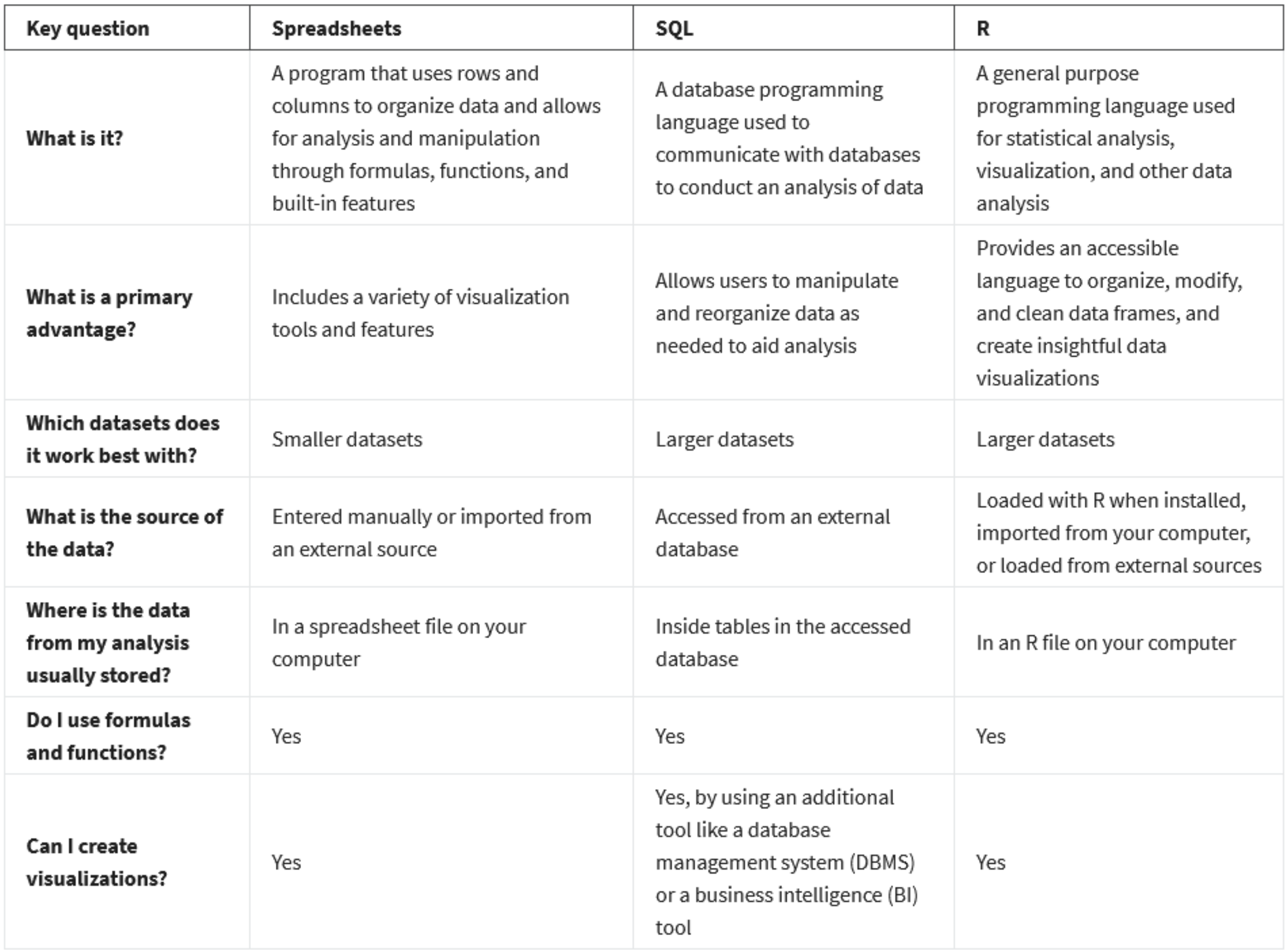
Languages	R	Python
Common features	- Open-source - Data stored in data frames - Formulas and functions readily available - Community for code development and support	- Open-source - Data stored in data frames - Formulas and functions readily available - Community for code development and support
Unique advantages	- Data manipulation, data visualization, and statistics packages - "Scalpel" approach to data: *find packages to do what you want with the data*	- Easy syntax for machine learning needs - Integrates with cloud platforms like Google Cloud, Amazon Web Services, and Azure
Unique challenges	- Inconsistent naming conventions make it harder for beginners to select the right functions - Methods for handling variables may be a little complex for beginners to understand	- Many more decisions for beginners to make about data input/output, structure, variables, packages, and objects - "Swiss army knife" approach to data: *figure out a way to do what you want with the data*

Additional Readings

- https://medium.com/analytics-and-data/r-vs-python-a-comprehensive-guide-for-data-professionals-321e8dead598
- https://www.dataquest.io/blog/python-vs-r/
- https://blog.rstudio.com/2019/12/17/r-vs-python-what-s-the-best-for-language-for-data-science/

Programming as a data analyst

From spreadsheets to SQL to R



Key question	Spreadsheets	SQL	R
What is it?	A program that uses rows and columns to organize data and allows for analysis and manipulation through formulas, functions, and built-in features	A database programming language used to communicate with databases to conduct an analysis of data	A general purpose programming language used for statistical analysis, visualization, and other data analysis
What is a primary advantage?	Includes a variety of visualization tools and features	Allows users to manipulate and reorganize data as needed to aid analysis	Provides an accessible language to organize, modify, and clean data frames, and create insightful data visualizations
Which datasets does it work best with?	Smaller datasets	Larger datasets	Larger datasets
What is the source of the data?	Entered manually or imported from an external source	Accessed from an external database	Loaded with R when installed, imported from your computer, or loaded from external sources
Where is the data from my analysis usually stored?	In a spreadsheet file on your computer	Inside tables in the accessed database	In an R file on your computer
Do I use formulas and functions?	Yes	Yes	Yes
Can I create visualizations?	Yes	Yes, by using an additional tool like a database management system (DBMS) or a business intelligence (BI) tool	Yes

R Packages

Palmer penguins

- https://allisonhorst.github.io/palmerpenguins/ To view: View(penguins)

Tidyverse

- https://www.tidyverse.org/

Week-2

Understand basic programming concepts

Programming fundamentals

Case sensitive

The basic concepts of R

- functions,
- comments,
- variables,
- data types,
- vectors, and
- pipes.

PRINT

- print()
- ?print()

COMMAND

- #

Variables

- a<- ”Dhamu”
- b<- 10

Some commands to know

- **typeof(a)**
- **is.integer(a)**

Vector

vector is a group of data elements of the **same type** stored in a sequence in R.

Eg: z<- c(23,45,67)

Pipe

A pipe is a tool in R for expressing a sequence of multiple operations.

Represented by %>%

Vectors and lists in R

Some commands

- *typeof(a)*
- *is.integer(a)*

*List **different data type***

- *list("a", 1L, 1.5, TRUE)*

Naming list

- *list('Chicago' = 1, 'New York' = 2, 'Los Angeles' = 3)*
- *https://r4ds.had.co.nz/vectors.html#vectors*

For more information refer pdf

- " *M7_W2_1_Vectors and lists in R.pdf* "

Dates and times in R

Install

- *install.packages("tidyverse")*

Load

- *library(tidyverse)*
- *library(lubridate)*

Then

- *today()*
- *now()*

Converting from strings

- *ymd("2021-01-20")*
- *mdy("January 20th, 2021")*
- *dmy("20-Jan-2021")*
- *ymd_hms("2021-01-20 20:11:59")*
- *mdy_hm("01/20/2021 08:01")*

Other common data structures

- data.frame(x = c(1, 2, 3) , y = c(1.5, 5.5, 7.5))
- dir.create ("destination_folder")
- file.create (“new_text_file.txt”)
- file.copy (“new_text_file.txt” , “destination_folder”)
- unlink (“some_.file.csv”) **to delete**
- matrix(c(3:8), nrow = 2)

Explore coding in R

Operators and calculations

Assignment operators

Assignment operators are used to assign values to variables and vectors.

Logical operators and conditional statements

Logical operators

- AND (sometimes represented as & or && in R)
- OR (sometimes represented as | or || in R)
- NOT (!)

Conditional statements

- if()
- else()
- else if()

For more information refer pdf

- “ M7_W2_2_Logical operators and conditional statements.pdf "

Learning about R packages

Packages in R

- Tidyverse   -It has inbuilt 8 packages

Pipe %>%

Available R packages

Choosing the right packages

- https://www.tidyverse.org/
- https://support.rstudio.com/hc/en-us/articles/201057987-Quick-list-of-useful-R-packages
- https://cran.r-project.org/web/views/

R resources for more help

- https://www.rstudio.com/
- https://blog.rstudio.com/
- https://blog.rstudio.com/categories/featured/
- https://stackoverflow.blog/
- https://www.r-bloggers.com/2015/12/how-to-learn-r-2/#h.y5b98o9o2h1r

Week-3

Explore data and R

Working with data frames

- *install.packages("tidyverse")*
- *library(ggplot2)*
- *View(diamonds)*
- ***head(diamonds), glimpse()***
- ***str(diamonds)*** *-some info included column names, data type etc.*
- *colnames(diamonds) -This is only for column names.*
- *mutate() -Make changes in the dataframe.*

Highlighted are summary functions

More about tibbles

Tibbles

Which are a super useful tool for organizing data in R.

- *as_tibble(diamonds)*

Data-import basics

Import data. and readxl Package

Refer to the pdf named as " **M7_W3_Data-import basics.pdf** "

Cleaning data

Cleaning up with the basics

Install

- install.packages(“here”) - Reading files
- install.packages(“skimr”) - Summarizing data
- install.packages(“janitor”) - Cleaning data
- install.packages(“dplyr”)

Load

- library(“here”)
- library(“skimr”)
- library(“janitor”)
- library(“dplyr”)

These are the packages required for data cleaning.

There's a few different functions that we can use to get summaries of our data frame.

- Skim without charts,
- glimpse,
- head, and
- Select.

Some functions

- rename() - To Rename the column
- rename_with() - Rename with upper case
- clean_names() - Make sure that the column names are unique and consistent.

File-naming conventions

Give easily understandable file name with underscores

- https://speakerdeck.com/jennybc/how-to-name-files

- https://libguides.princeton.edu/c.php?g=102546&p=930626#:~:text=File%20naming%20best%20practices%3A&text=File%20names%20should%20be%20short,date%20format%20ISO%208601%3A%20YYYYMMDD

More on R operators

In R, there are four main types of operators:

1. Arithmetic
2. Relational
3. Logical
4. Assignment

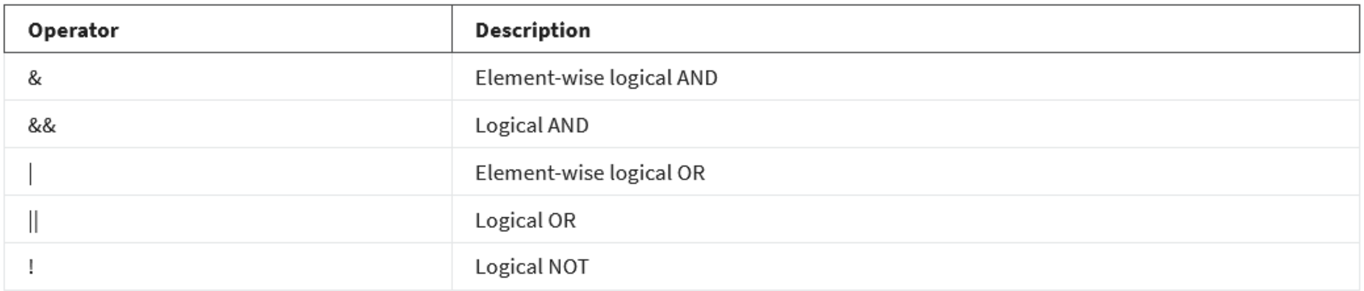
Arithmetic operator

Operator	Description	Example Code	Result/ Output
+	Addition	x + y	[1] 7
-	Subtraction	x - y	[1] -3
*	Multiplication	x * y	[1] 10
/	Division	x / y	[1] 0.4
%%	Modulus (returns the remainder after division)	y %% x	[1] 1
%/%	Integer division (returns an integer value after division)	y%/% x	[1] 2
^	Exponent	y ^ x	[1]25

Relational operators

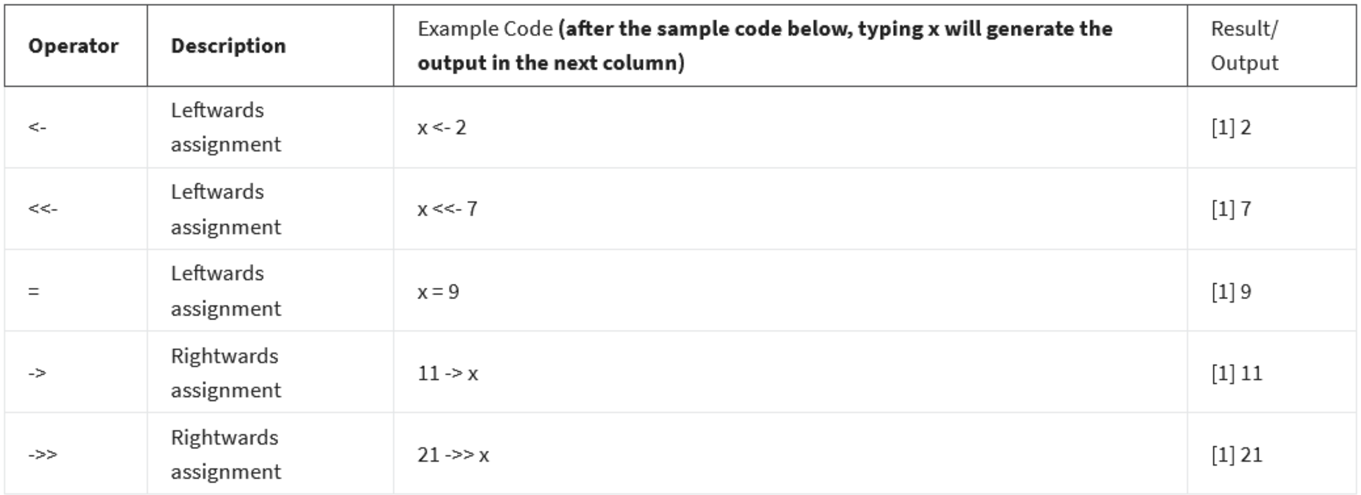
Operator	**Description**	Example Code	Result/Output
<	Less than	x < y	[1] TRUE
>	Greater than	x > y	[1] FALSE
<=	Less than or equal to	x < = 2	[1] TRUE
>=	Greater than or equal to	y >= 10	[1] FALSE
==	Equal to	y == 5	[1] TRUE
!=	Not equal to	x != 2	[1] FALSE

Logical operators



Operator	Description
&	Element-wise logical AND
&&	Logical AND
|	Element-wise logical OR
||	Logical OR
!	Logical NOT

Assignment operators



Operator	Description	Example Code **(after the sample code below, typing x will generate the output in the next column)**	Result/ Output
<-	Leftwards assignment	x <- 2	[1] 2
<<-	Leftwards assignment	x <<- 7	[1] 7
=	Leftwards assignment	x = 9	[1] 9
->	Rightwards assignment	11 -> x	[1] 11
->>	Rightwards assignment	21 ->> x	[1] 21

Organize your data

Some functions to organize the data.

It will be helpful to turn information into knowledge.

- *arrange()* - *Sorting*
- *group_by()*
- *filter()*

Transforming data

Some function to transform data

- Separate()
- unite()
- mutate()

Wide to long with tidyr

Additional resources

- https://tidyr.tidyverse.org/articles/pivot.html
- https://www.tidyverse.org/
- https://rladiessydney.org/courses/ryouwithme/02-cleanitup-5/
- https://scc.ms.unimelb.edu.au/resources-list/simple-r-scripts-for-analysis/r-scripts

Take a closer look at the data

Same data, different outcome

Anscombe's quartet has four datasets that have nearly identical summary statistics.

The bias function

bias()

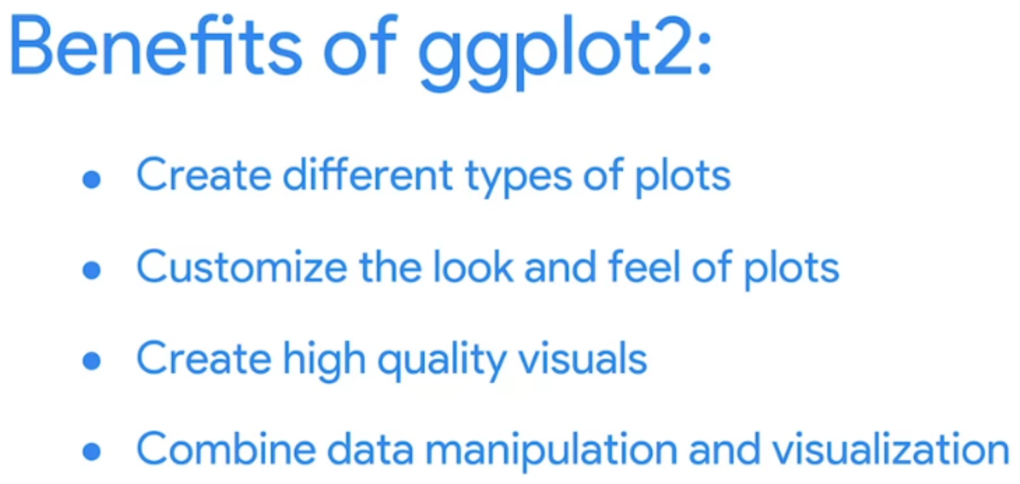
Working with biased data

- https://www.rdocumentation.org/packages/SimDesign/versions/2.2/topics/bias
- https://datasciencebox.org/ethics.html

Week-4

Create data visualizations in R

Visualization basics in R and tidyverse



Benefits of ggplot2:

- Create different types of plots
- Customize the look and feel of plots
- Create high quality visuals
- Combine data manipulation and visualization

some core concepts in ggplot2:

- *aesthetics,*
- *geoms,*
- *facets,*
- *labels and annotations.*

Facets

Facets let you display smaller groups or subsets of your data.
With facets, you can create separate plots for all the variables in your dataset.

Common problems when visualizing in R

- *Check the pdf "M7_W4_Common problems when visualizing in R.pdf"*

Getting started with ggplot()

- ggplot() in R

Explore aesthetics in analysis

Aesthetic attributes

There are three aesthetic attributes in ggplot2:

- Color: this allows you to change the color of all of the points on your plot, or the color of each data group
- Size: this allows you to change the size of the points on your plot by data group
- Shape: this allows you to change the shape of the points on your plot by data group

Additional resources

- https://ggplot2.tidyverse.org/
- http://statseducation.com/Introduction-to-R/modules/graphics/aesthetics/
- https://www.rdocumentation.org/packages/ggplot2/versions/3.3.3/topics/aes

Smoothing

Smoothing enables the detection of a data trend even when you can't easily notice a trend from the plotted data points.

Two types of smoothing

Loess smoothing

The loess smoothing process is best for smoothing plots with less than 1000 points.

Gam smoothing

Gam smoothing, or generalized additive model smoothing, is useful for smoothing plots with a large number of points.

Filtering and plots

Annotate and save visualizations

Drawing arrows and shapes in R

- https://ggplot2.tidyverse.org/reference/annotate.html
- https://www.r-graph-gallery.com/233-add-annotations-on-ggplot2-chart.html
- https://ggplot2-book.org/annotations.html
- https://www.r-bloggers.com/2017/02/how-to-annotate-a-plot-in-ggplot2/
- https://viz-ggplot2.rsquaredacademy.com/ggplot2-text-annotations.html

Saving images without ggsave()

- https://ggplot2.tidyverse.org/reference/ggsave.html#saving-images-without-ggsave-
- https://www.tidyverse.org/
- https://www.datanovia.com/en/blog/how-to-save-a-ggplot/
- https://www.datamentor.io/r-programming/saving-plot/

Week-5

Develop documentation and reports

R Markdown resources

R Markdown documentation

- https://rmarkdown.rstudio.com/lesson-1.html

R Markdown reference materials

- https://rmarkdown.rstudio.com/lesson-15.html
- https://www.rstudio.com/wp-content/uploads/2015/03/rmarkdown-reference.pdf?_ga=2.49295910.1034302809.1602760608-739985330.1601281773

R for Data Science book

- https://r4ds.had.co.nz/communicate-intro.html

R Markdown: The Definitive Guide

- https://bookdown.org/yihui/rmarkdown/
- https://bookdown.org/yihui/rmarkdown/installation.html
- https://bookdown.org/yihui/rmarkdown/documents.html
- https://bookdown.org/yihui/rmarkdown/dashboards.html
- https://bookdown.org/yihui/rmarkdown/parameterized-reports.html

Optional: Jupyter notebooks

- https://colab.research.google.com/notebooks/intro.ipynb
- https://www.kaggle.com/docs/notebooks
- https://jupyter.org/
- https://realpython.com/jupyter-notebook-introduction/

To learn about basic formatting in Jupyter notebooks

- https://jupyter-notebook.readthedocs.io/en/stable/notebook.html
- https://gtribello.github.io/mathNET/assets/notebook-writing.html
- https://medium.com/analytics-vidhya/the-jupyter-notebook-formatting-guide-873ab39f765e

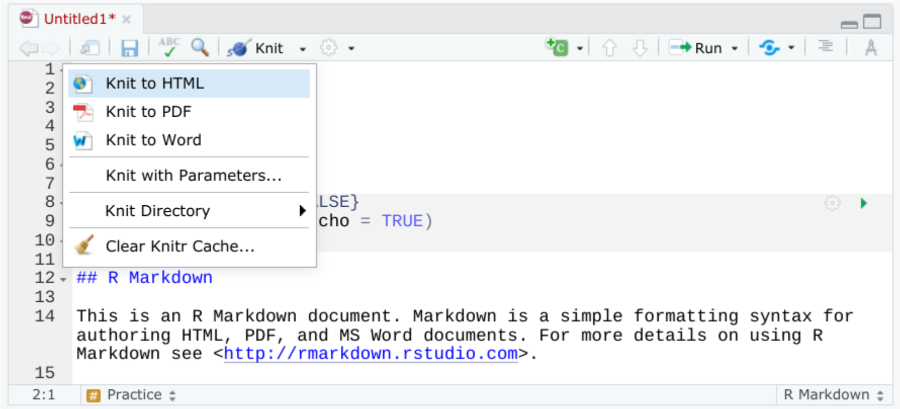
Understand code chunks and exports

Adding code chunks to R Markdown notebooks

Output formats in R Markdown

- Refer pdf " M7_W5_Output formats in R Markdown.pdf "

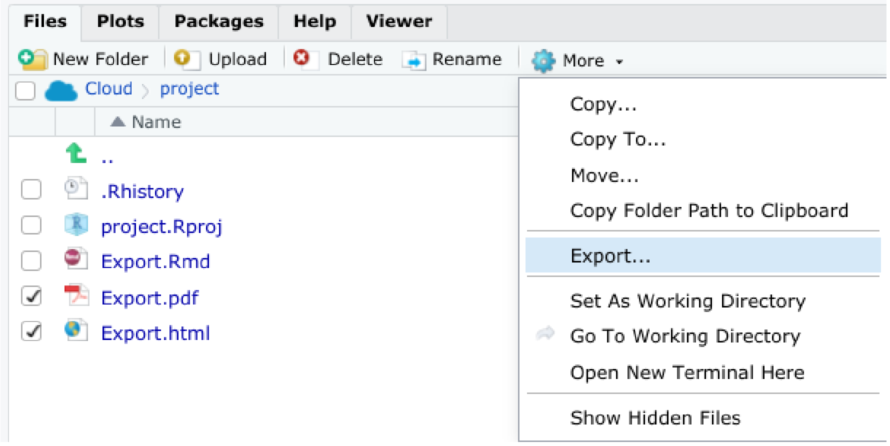
Exporting your R Markdown notebook



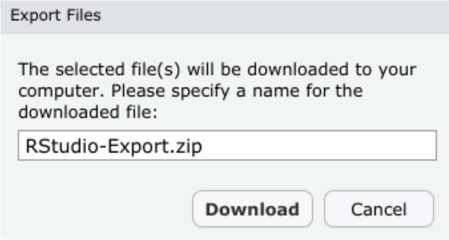
```
LSE}
cho = TRUE)

## R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for
authoring HTML, PDF, and MS Word documents. For more details on using R
Markdown see <http://rmarkdown.rstudio.com>.
```



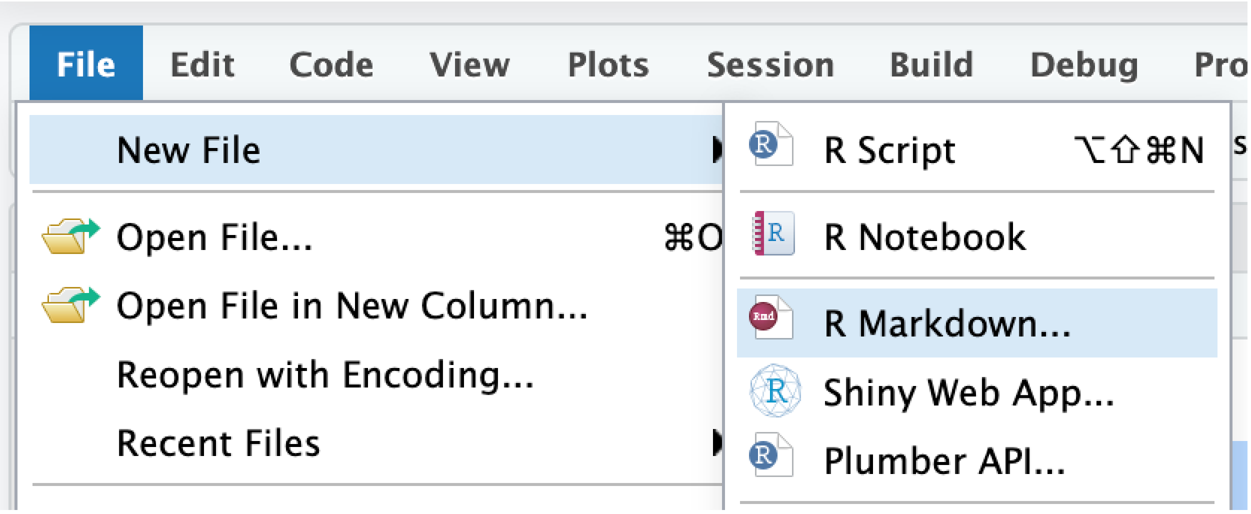
- □ .Rhistory
- □ project.Rproj
- □ Export.Rmd
- ☑ Export.pdf
- ☑ Export.html

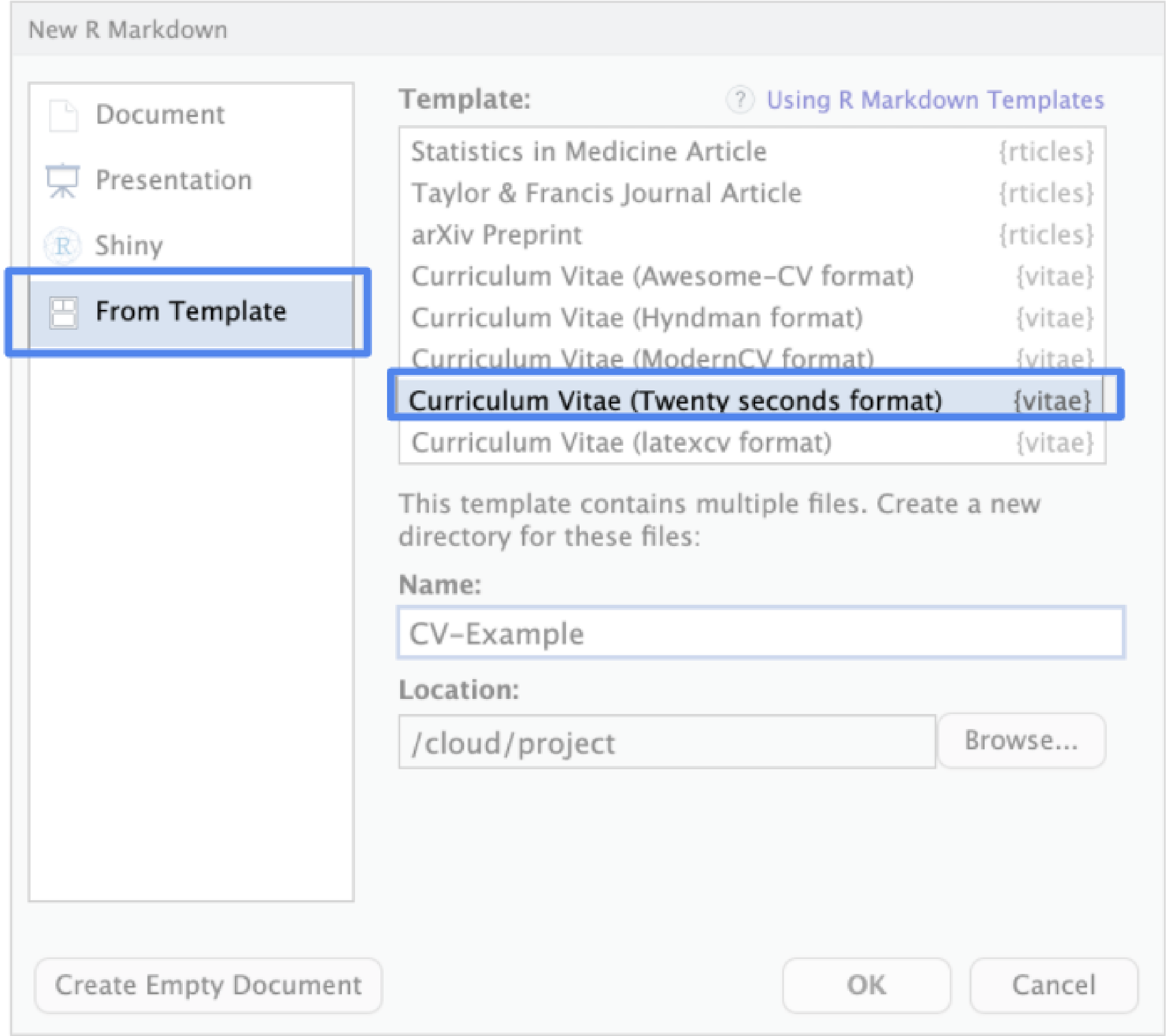


The selected file(s) will be downloaded to your computer. Please specify a name for the downloaded file:

RStudio-Export.zip

Using R Markdown templates





Template:

Statistics in Medicine Article	{rticles}
Taylor & Francis Journal Article	{rticles}
arXiv Preprint	{rticles}
Curriculum Vitae (Awesome-CV format)	{vitae}
Curriculum Vitae (Hyndman format)	{vitae}
Curriculum Vitae (ModernCV format)	{vitae}
Curriculum Vitae (Twenty seconds format)	{vitae}
Curriculum Vitae (latexcv format)	{vitae}

This template contains multiple files. Create a new directory for these files:

Name: CV-Example

Location: /cloud/project

Quick Review

Week-1

- *Introduction to R programming language*

Week-2

- *Basic concepts*
- *R Packages*

Week-3

- *Data frame*
- *Cleaning data*
- *Checking for biasing*

Week-4

- *ggplot()*
- *Save plotted images*

Week-5

- *Jupyter notebook*
- *R Markdown notebook*

Dhamodharan
30/10/2021