1. 导入数据
2. Absolute paths

从 / 开始，从最开始的文件夹依次往下，中间用 / 。

Relative paths

有一个 working directory, 从下级文件夹开始找，中间用 /

1. **Reading tabular data from a plain text file into R**
   1. 如果文本中间用 ， 隔开，使用read\_csv

Ex. canlang\_data <- read\_csv("data/can\_lang.csv")

* 1. 如果前几列是关于数据说明，我们需要skip前几列一直到有数据表头的那一列

文本

低可信度描述已自动生成Ex.

* 1. 中间用 tab（空格）隔开，使用read\_tsv

图片包含 徽标

描述已自动生成Ex.

* 1. General way: read\_delim

将文件作为第一个参数，制表符作为第二个参数（例如\t or ,），false作为第三个参数说明数据里没有column name （通常情况下默认true）

图片包含 文本

描述已自动生成Ex.

随后，我们想要他有column name，需要把默认的x1 x2 x3替换，使用rename

图形用户界面, 文本, 应用程序

描述已自动生成Ex.

* 1. 从URL 读取数据（远程文件，非计算机上的）

图片包含 文本

描述已自动生成Ex

* 1. 当U RL 的数据不规范的时候，我们使用download.file 函数先下载到计算机在读取。第一个参数是 URL ，第二个参数是希望下载路径

文本

中度可信度描述已自动生成Ex

* 1. **Reading tabular data from a Microsoft Excel file （.xlsx）**

这种文件和csv不同。Csv是纯文本文件，而.xlsx不是

使用 read\_excel 读取文件

图形用户界面

低可信度描述已自动生成Ex

* 1. Summary of read\_\* functions

图形用户界面, 文本, 应用程序, 电子邮件

描述已自动生成

**Reading data from a database**

1. Reading data from a SQLite database

这些文件通常是(.db) or (.sqlite) 扩展名

使用 DBI package, dbConnect来建立R和数据库的通道，但不会直接读取数据

图片包含 文本

描述已自动生成Ex

* 1. 数据库通常有多个表，需要dbListTables 函数获取所有表名称

文本

中度可信度描述已自动生成Ex

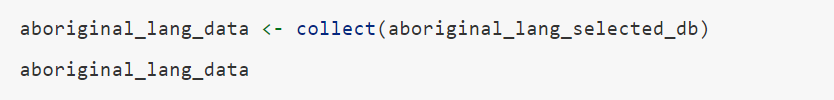
上图说明只有一个表

* 1. 使用 dbplyr package, tbl function 可以引用该表（仍未获取或存储），允许我们进行更多操作（如select，etc）

图片包含 文本

描述已自动生成Ex

* 1. 当完成select等对数据的操作之后，（确保进行筛选后再导入，否则数据会很庞大导致计算机崩溃）。我们可以使用collect 函数把数据存储导入

Ex

1. **Reading data from a PostgreSQL database**

是通过网络访问数据，我们需要 RPostgres package， 还需要诸多参数

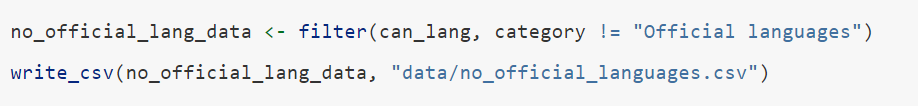
文本

描述已自动生成ex.

建立通道之后，剩下的步骤和SQLite database相同

1. 建立数据库的优势，允许多个用户访问，存储庞大数据集
2. 把数据从R中导入到.csv文件

使用 tidyverse package 中的 write\_csv function

Ex.

二、 Wrangling Data

表格

描述已自动生成1. data frame （数据框，类似表格）:

Variable: 可测量的特征

Observation: 针对某个实体的所有测量数据

Value: 针对某个实体单个测量值

图形用户界面, 文本, 应用程序, 电子邮件

描述已自动生成2. vector：包含一个或多个元素，属于相同数据类型。Eg. Region, year, population都是vector

图示

描述已自动生成

3. list ：包含的元素可以属于不同的数据类型。

4. 一种特殊数据类型：tibble， 也是data frames，

可以用class判断。Eg. Class(can\_lang)

* 1. 如果数据对象是普通数据框，class 会返回 "data.frame"。
  2. 如果数据对象是 tibble，class 会返回 ["tbl\_df", "tbl", "data.frame"]。

图形用户界面, 文本, 应用程序, 电子邮件

描述已自动生成5. summary:

6. tidy data: ①每一行（row）代表一个独立的观测值（observation）

②每一列（column）代表一个独立的变量（variable）

③每个数值（value）只占据单元格中的一个唯一位置，即不会与其他值共享同一个单元格

7. how to make a tidy data:

①从宽变长

图形用户界面, 应用程序

描述已自动生成

②从长变宽

图片包含 表格

描述已自动生成

③当单元格内不是单个value

日程表

中度可信度描述已自动生成

当数据类型错误时（如most\_at\_home is chr）, 可以加一行如下

图形用户界面, 文本, 应用程序

描述已自动生成

8. select

文本

描述已自动生成Eg.

Some easier methods:

① 从哪列到哪列

图片包含 图示

描述已自动生成

②以什么开头

图片包含 图表

描述已自动生成

③ 包含什么

图片包含 文本

描述已自动生成

9. filter

日程表

描述已自动生成

② 除了该类

图片包含 图表

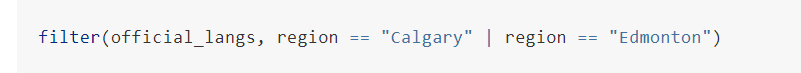
描述已自动生成

③

多个要求



④两个条件满足一个（or）



⑤

使用%in%来检查两个vector是不是一样的。和==的区别是==会一项一项比对，而%in%会看第一项在第二项里有没有，i.e.看两项包含的东西一样而不是顺序一样

图形用户界面, 应用程序

描述已自动生成

⑥

大于小于

图片包含 图表

描述已自动生成

10. c (链接)

文本

描述已自动生成

11. mutate

图形用户界面, 文本, 应用程序

描述已自动生成① 我们想直接更换数据类型，可以使用如下function。由于使用了和原来列名相同的列名，因此创建出来的新列会直接覆盖旧列。因此我们就直接实现了数据格式的转换

②mutate出来一个新列

图形用户界面, 文本, 应用程序

描述已自动生成

Notes： 这里的city\_pops是刚才使用c function弄出来的，所以新列会一项一项的除。因此需要保证顺序一致（按位置匹配）

12. pipe operator, |>

图形用户界面, 文本, 应用程序, Word

描述已自动生成

Another way is (其实都一样，会自动把上一行的output给下一行当数据)

图形用户界面, 文本, 应用程序

描述已自动生成

13. summarize：第一个参数是数据，后边的是我们想要执行的操作。

① 总结出特定语言作为家庭主要语言的加拿大人数量的**最小值和最大值。**

图形用户界面, 文本, 应用程序

描述已自动生成

②当数据中有NA（无数据）而我们想忽略他看其他值时

文本

描述已自动生成

③ groupby：可以先把数据分成许多组，然后对每个组进行相同的summarize，最后再结合起来就可以。（第一个参数是数据，第二个参数是要分组的列）

图形用户界面, 文本, 应用程序

描述已自动生成

④ 对每一列都进行操作：summarize + across

图形用户界面, 文本, 应用程序

描述已自动生成

当数据中有NA的时候，

图表

低可信度描述已自动生成

⑤对每一列都进行操作：map，会将函数应用到每一列，因此或许我们需要一个select来选出来我们希望应用到的列

徽标

中度可信度描述已自动生成

然而有时出来的不是data frame形式，我们或许想要另一种形式

图形用户界面, 文本, 应用程序, 电子邮件

描述已自动生成

当有NA value时

文本

中度可信度描述已自动生成

14. 对多个列应用某个函数

图片包含 图表

描述已自动生成

As.integer is the function

15. 对每一行的多个列进行相同的函数

Eg. 想取出每一行的最大值

表格

描述已自动生成

在mutate前面加一个rowwise

16. choose the correct graph type

文本, 信件

描述已自动生成

17. 基本代码：

图形用户界面, 文本

描述已自动生成

And geom\_point, geom\_bar

Theme函数中的text用于设定字体大小

18. 使用xlim缩放x轴范围，使用ylim缩放y轴范围。Date来自lubridate package，可以将字符串转为日期。

Eg. Focus on 1990 – 1994

图表, 折线图

描述已自动生成

19. 使用\n 把xlab换行（ylab换行）

20. 当散点图左下角和右上角（极小值和极大值）特别聚集的时候，我们可以对x轴y轴都进行对数缩放，并使用label\_comma给x坐标和y坐标千位加逗号以便阅读

图表, 散点图

描述已自动生成

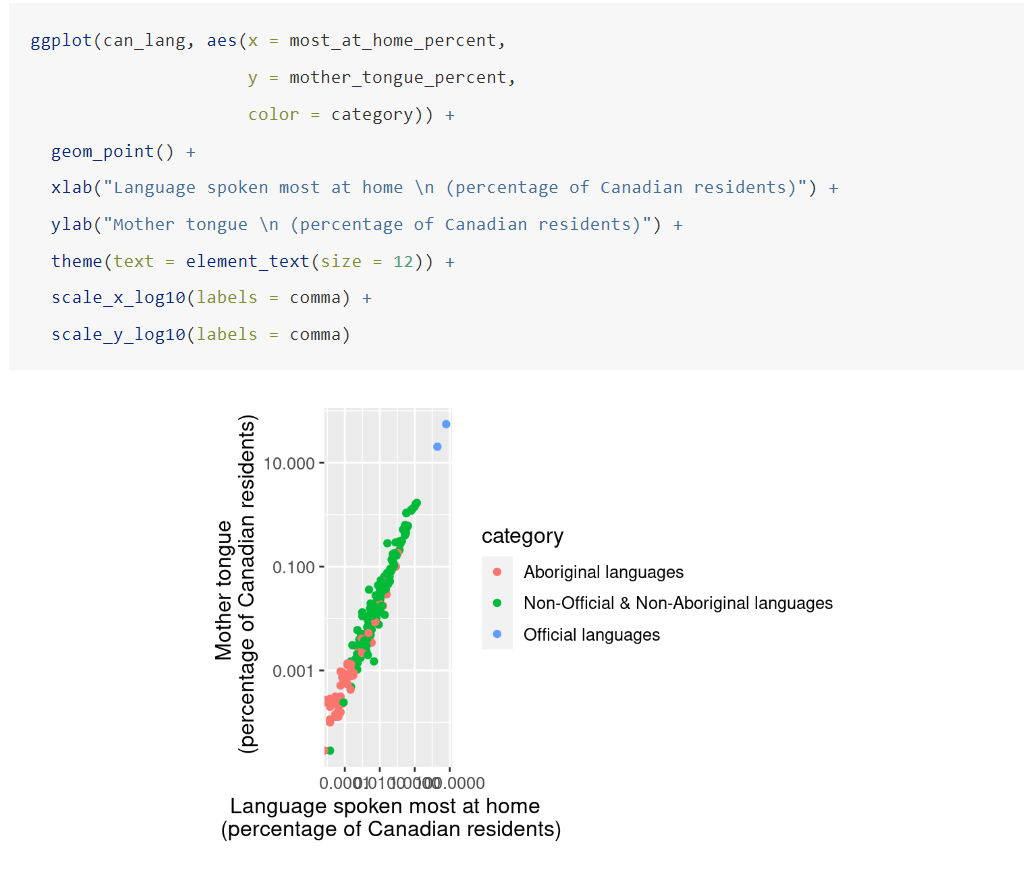
然后，为了方便理解，我们可以使用mutate计算出母语使用者比例和家庭语言使用者的比例，放在新列，再select出这两列，再跟上面一样进行绘图

21. how to describe the relationship between two variables

文本

描述已自动生成

22. 着色：在aes中添加参数指定某个列用于着色点



上图图例占用了太多位置，改变图例位置

文本

描述已自动生成

图形用户界面, 文本, 应用程序

描述已自动生成Eg

进一步优化我们的图像：可以根据种类区分形状，以及选择调色包

使用display.brewer.all 调用调色板

图形用户界面

描述已自动生成

23. bar plot：特别注意的是该函数的y值默认计数，如果希望计算特定东西需要identity

图表

低可信度描述已自动生成

我们只想关注于面积前12的土地，并且xlab重合在一起无法阅读，因此我们使用slice\_max筛选出前十二的土地，并且将该图横向（就是交换x坐标和y坐标）

图片包含 日程表

描述已自动生成

进一步优化：着色和排序

图表

描述已自动生成

文本, 应用程序

描述已自动生成

文本

中度可信度描述已自动生成

24. histograms （geom\_histogram）

只需设定x坐标，可以添加虚线(geom\_vline)画线, (linetype = “dashed”)来变成虚线。可以比较数值。Position = identity 确保柱状图不重叠， alpha让柱子透明，fill = expt代表要根据expt变量着色x

图表, 直方图

描述已自动生成

注意，这里仍然是一种颜色，原因是Expt是整数variable，而不是类别。因此我们需要使用as\_factor (Expt)对这些整数进行分类

文本

描述已自动生成

图表, 条形图, 直方图

描述已自动生成

上图仍会造成混淆，因此我们来创建子图

图片包含 图形用户界面

描述已自动生成

如果需要按行拆分绘图，用rows，用cols区分是水平排列两张图还是竖直排列两张图

最后优化：我们只知道速度，但不知道关于光速的偏差度。我们mutate一个新cols，并且再创建标签。

图片包含 图表

描述已自动生成

Binwidth 0.01 最合适，0.1过宽 0.001过窄

25. ggplot允许我们总是可以使用+来增加layers

Eg. 我们想给上图加一个title

图表

描述已自动生成

描述数据方法：见textbook4.6

27. 保存图像方法

表格

描述已自动生成

Vector画质更高，放大无损，但是加载更慢。