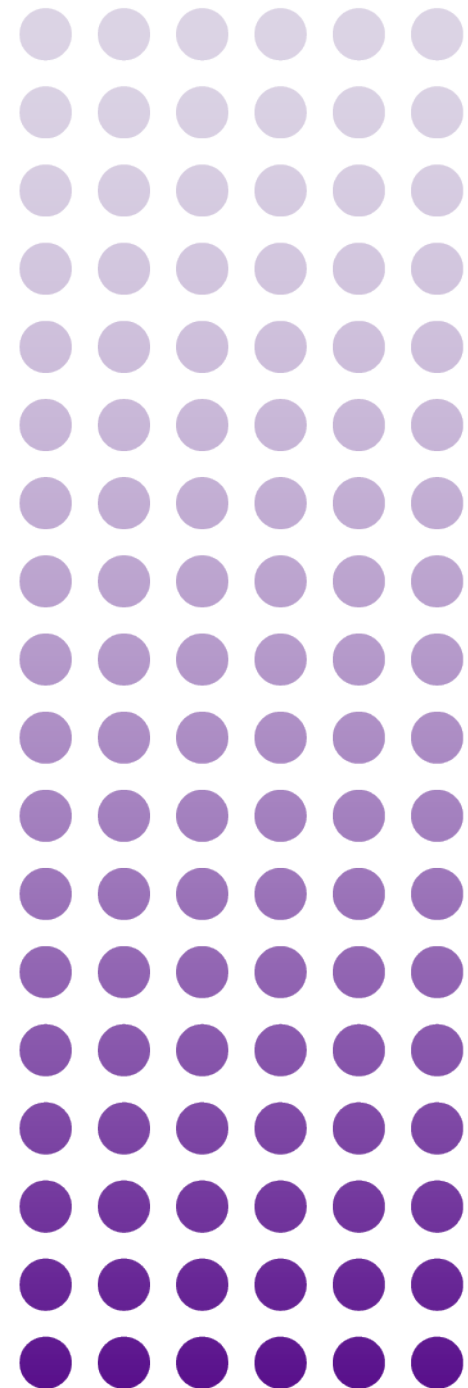


Reproducibility Workshop: Part 2

R Markdown for Reproducible Data Processing and Analysis

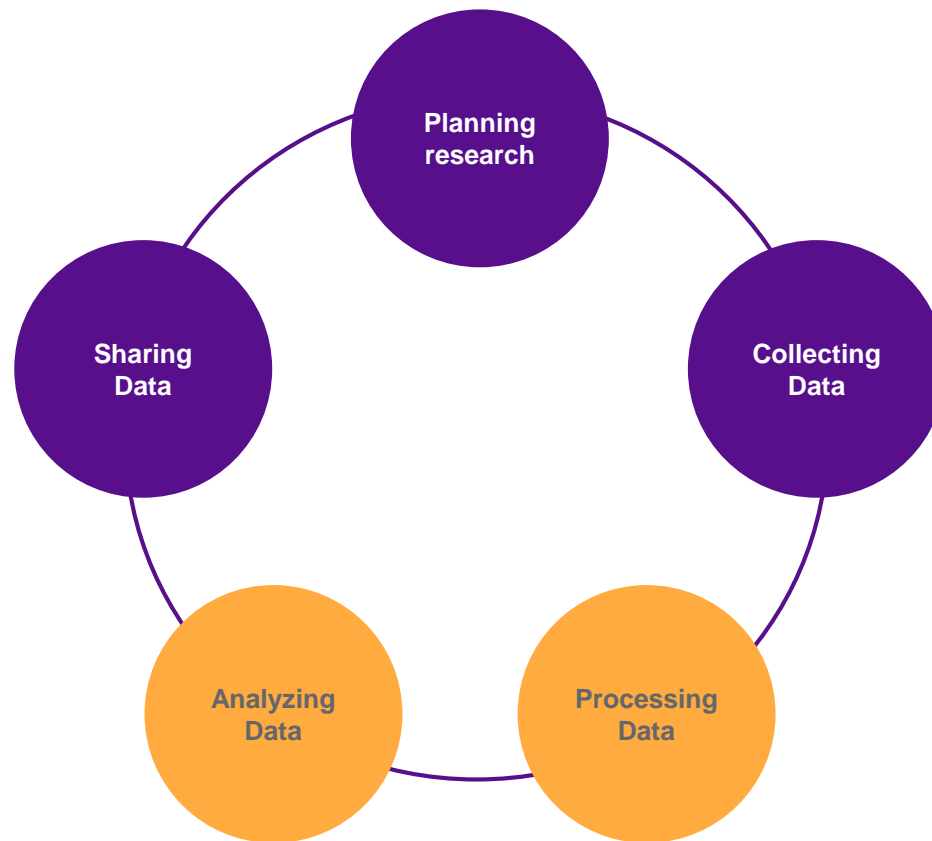
Alisa Surkis, PhD, MLS
Assistant Director, Research Data and Metrics
NYU Health Sciences Library



Class Overview

- Why use R?
- Why use R Markdown for Reproducibility?
- Rstudio
- Terminology and Syntax
- Setting up an R Project
- R Basics
- Reproducible Case Study: Data Processing and Analysis

Data Lifecycle



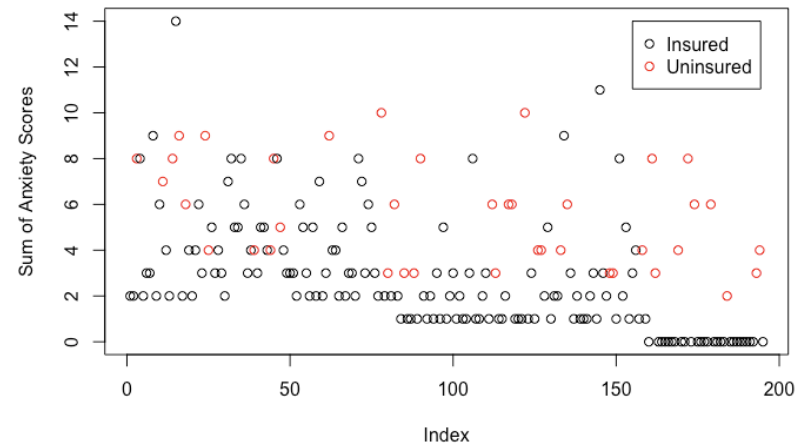
Why use R?

```
notUninsured <- select(filter(AnalysisData, ever_uninsured == 0), anxietySum, difficultyFunctioning)
Uninsured <- select(filter(AnalysisData, ever_uninsured == 1), anxietySum, difficultyFunctioning)
diffAnxiety <- t.test(notUninsured$anxietySum, Uninsured$anxietySum)
diffAnxietyHyp <- ifelse(diffAnxiety$p.value < 0.05, "Rejected", "Accepted")
```

```
EHRData$DATE <- mdy(EHRData$DATE)
```

```
EHRData <- spread(EHRData, key = description, value=value)
```

```
EHRData <- select(EHRData, patient_id="PATIENT", date_of_visit = "DATE", ir  
housing_status = "Housing status", HIV_status = "HIV status in {nominal}")
```

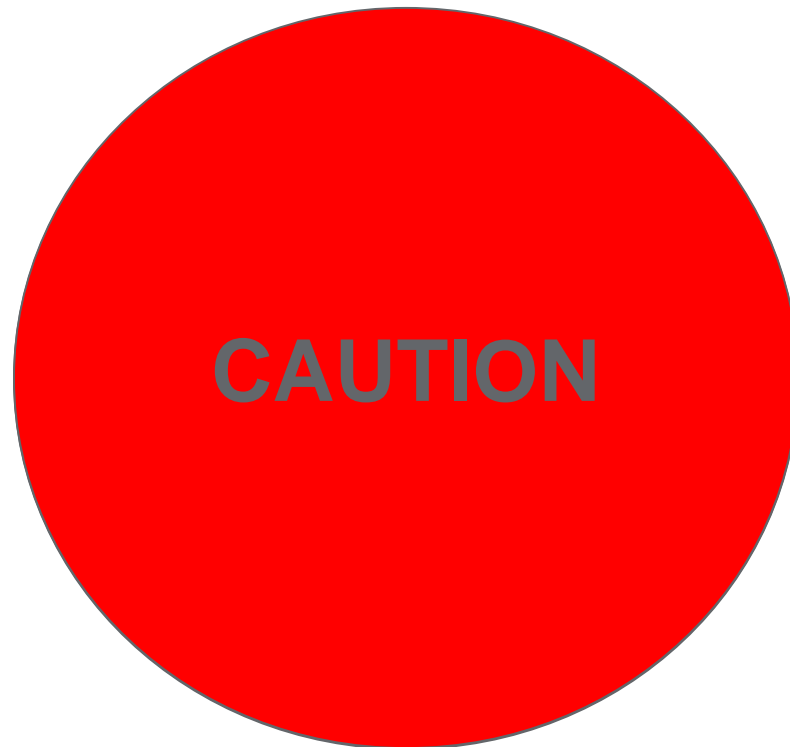


Everything in one place: data processing, analysis, and visualization

Why R?

It's free!

Large user community contributing functionality



Why R Markdown?

Creates formatted documents that integrate:

- Code -- HOW
- Narrative Text -- WHY
- Output -- WHAT

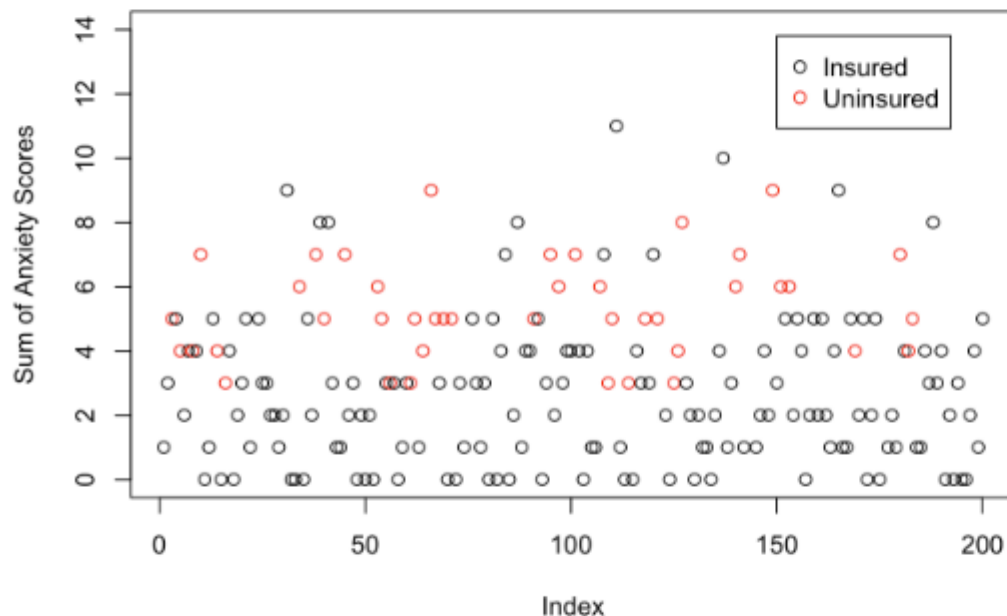
Produces an analysis notebook

```
analysis$sex <- as.factor(analysis$sex)
```

Uninsured analysis

Plot the anxietySum with symbol color determined by if ever uninsured Separate the insured and uninsured and run a t-test

```
plot(analysis$anxietySum, col=analysis$ever_uninsured,ylab="Sum of Anxiety Scores",ylim=c(0,14))  
legend(150,13.8,legend=c("Insured","Uninsured"),pch=1,col=1:2)
```



```
notUninsured <- filter(analysis, ever_uninsured == FALSE)  
Uninsured <- filter(analysis, ever_uninsured == TRUE)  
  
diffAnxiety <- t.test(notUninsured$anxietySum, Uninsured$anxietySum)  
diffAnxietyHyp <- ifelse(diffAnxiety$p.value < 0.05, "Rejected","Accepted")
```

Results of analysis with uninsured vs insured populations

The hypothesis that there is no difference between the insured and uninsured populations in their total anxiety is Rejected with p-value of $9.038641710 \times 10^{-14}$

RStudio

The screenshot displays the RStudio IDE with the following components:

- Source Pane:** Contains an R script named `PullAllICTSAPubs.R`. The script performs the following steps:
 - 1. Sets the working directory to `~/Google Drive/work/Publication Classification/Analysis/R Code/`.
 - 2. Loads the `library(rentrez)` and `library(XML)`.
 - 3. Reads a CSV file `../All CTSA Grant Numbers.csv` into `GrantNum`, with `stringsAsFactors = FALSE` and `na.strings=""`.
 - 4. Extracts `UL1`, `TL1`, and `KL2` columns from `GrantNum`.
 - 5. Creates a single vector `AllGrantNums` by appending `UL1`, `TL1`, and `KL2`.
 - 6. Extracts the "kernel" (characters 5-13) from `AllGrantNums` into `kernels`.
 - 7. Creates a full grant number vector `fullNums` by prepending a space to `kernels`.
 - 8. Searches PubMed for variants of `fullNums` using `entrez_search` (term=`x`, retmax=5000).
 - 9. Extracts `pmids` from the search results.
 - 10. Removes `pmids` with no abstracts.
- Environment Pane:** Shows the `GrantNum` data frame with 178 observations and 6 variables. The variables are `AllGrantNums`, `KL2`, `TL1`, and `UL1`, all of type `chr`.
- Console Pane:** Displays the R version (3.4.1), copyright information, and workspace status. It also shows the execution of the first few lines of the script.
- Help Pane:** Displays the documentation for the `update.packages` function, including a description, usage, and examples.

RStudio

The screenshot displays the RStudio interface with four main panes:

- Source Pane:** Contains an R script named `PullAllICTSAPubs.R`. The script reads a CSV file, processes grant numbers, and uses `entrez_search` to fetch PubMed data. Lines 8-11 and 13-14 are highlighted in blue.
- Environment Pane:** Shows the current environment with 178 observations and 6 variables. The `Values` section displays the first few rows of the data frame.
- Console Pane:** Displays the R version (3.4.1), copyright information, and a list of contributors. A yellow box with the word **Console** is overlaid on this pane.
- Files/Plots/Packages/Help/Viewer Pane:** The **Help** tab is active, showing the documentation for the `update.packages` function. The title is "Compare Installed Packages with CRAN-like Repositories".

Environment Pane Data:

GrantNum	178 obs. of 6 variables
AllGrantNums	chr [1:495] "UL1 TR000086" "UL1 RR025750" "UL1 TR001073" "UL1 TR000157" "U...
KL2	chr [1:178] "KL2 TR000088" "KL2 RR025749" "KL2 TR001071" "KL2 TR000158" "K...
TL1	chr [1:139] "TL1 TR000087" "TL1 RR025748" "TL1 TR001072" "TL1 TR000159" "T...
UL1	chr [1:178] "UL1 TR000086" "UL1 RR025750" "UL1 TR001073" "UL1 TR000157" "U...

Console Output:

```
R version 3.4.1 (2017-06-30) -- "Single Candle"
Copyright (C) 2017 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.start()' for an HTML browser interface to help,
Type 'q()' to quit R.

[workspace loaded from ~/.RData]

> setwd("~/Users/surkia01/Google Drive/Work/Publication Classification/Analysis/R Code")
> library(rentrez)
> 
> GrantNum <- read.csv("../All CTSA Grant Numbers.csv", stringsAsFactors = FALSE, na.strings="")
> UL1 <- GrantNum$UL1[!is.na(GrantNum$UL1)]
> TL1 <- GrantNum$TL1[!is.na(GrantNum$TL1)]
> KL2 <- GrantNum$KL2[!is.na(GrantNum$KL2)]
> 
> ## create a single vector with all grant numbers listed
> AllGrantNums <- as.vector(append(append(UL1,TL1),KL2))
> 
> |
```

Help Pane Content:

Compare Installed Packages with CRAN-like Repositories

Description

`old.packages` indicates packages which have a (suitable) later version on the repositories whereas `update.packages` offers to download and install such packages.

`new.packages` looks for (suitable) packages on the repositories that are not already installed, and optionally offers them for installation.

Usage

```
update.packages(lib.loc = NULL, repos = getOption("repos"),
  contriburl = contrib.url(repos, type),
  method, instlib = NULL,
  ask = TRUE, available = NULL,
  oldPkgs = NULL, ..., checkBuilt = FALSE,
  type = getOption("pkgType"))

old.packages(lib.loc = NULL, repos = getOption("repos"),
  contriburl = contrib.url(repos, type),
  instPkgs = installed.packages(lib.loc = lib.loc),
  method, available = NULL, checkBuilt = FALSE,
  type = getOption("pkgType"))

new.packages(lib.loc = NULL, repos = getOption("repos"),
  contriburl = contrib.url(repos, type),
  instPkgs = installed.packages(lib.loc = lib.loc),
  method, available = NULL, ask = FALSE, ...)
```

RStudio

The screenshot displays the RStudio interface with three main panes:

- R Script Pane:** Contains R code for reading and processing grant data. A yellow box highlights the text "R Script Pane".
- Console:** Shows the R version (3.4.1) and the execution of the code from the script pane. A yellow box highlights the text "Console".
- Environment/History Pane:** Displays the current environment with a data object named "GrantNum" containing 178 observations of 6 variables. A table of values is shown below.

Environment/History Pane Data:

GrantNum	178 obs. of 6 variables
AllGrantNums	chr [1:495] "UL1 TR000086" "UL1 RR025750" "UL1 TR001073" "UL1 TR000157" "U...
KL2	chr [1:178] "KL2 TR000088" "KL2 RR025749" "KL2 TR001071" "KL2 TR000158" "K...
TL1	chr [1:139] "TL1 TR000087" "TL1 RR025748" "TL1 TR001072" "TL1 TR000159" "T...
UL1	chr [1:178] "UL1 TR000086" "UL1 RR025750" "UL1 TR001073" "UL1 TR000157" "U...

Console Output:

```
R version 3.4.1 (2017-06-30) -- "Single Candle"
Copyright (C) 2017 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.start()' for an HTML browser interface to help,
Type 'q()' to quit R.

[workspace loaded from ~/.RData]

> setwd("~/Google Drive/work/Publication Classification/Analysis/R Code")
> library(rentrez)
> 
> GrantNum <- read.csv("../All CTSA Grant Numbers.csv", stringsAsFactors = FALSE, na.strings="")
> UL1 <- GrantNum$UL1[!is.na(GrantNum$UL1)]
> TL1 <- GrantNum$TL1[!is.na(GrantNum$TL1)]
> KL2 <- GrantNum$KL2[!is.na(GrantNum$KL2)]
> 
> ## create a single vector with all grant numbers listed
> AllGrantNums <- as.vector(append(append(UL1,TL1),KL2))
> 
> |
```

RStudio

The image shows the RStudio interface with three panes highlighted by yellow boxes and labels:

- R Script Pane:** Displays the source code for a script named `PullAllICTSPubs.R`. The code includes comments and R commands for loading libraries, reading CSV files, and processing grant numbers.
- Environment Pane:** Shows the current environment with 178 observations and 6 variables. The variables listed are `GrantNum`, `AllGrantNums`, `KL2`, `TL1`, and `UL1`.
- Console:** Displays the output of the R script, including the R version (3.4.1), copyright information, and the execution of the script commands.

The bottom right pane shows the `update.packages` function documentation, which compares installed packages with CRAN-like repositories.

RStudio

The image shows the RStudio interface with four panes highlighted by yellow boxes:

- R Script Pane:** Contains R code for reading a CSV file, creating vectors, and applying functions. The code is as follows:

```
1 setwd("~/Google Drive/work/Publication Classification/Analysis/R Code")
2
3 ## load library to run entrez calls from R
4 library(rentrez)
5 library(xsl)
6
7 ## read in all CTSA grant nums and pull UL1, TL1, and KL2 grant nums
8 GrantNum <- read.csv("~/Google Drive/work/Publication Classification/Analysis/R Code/GrantNums.csv", na.strings="")
9 UL1 <- GrantNum$UL1
10 TL1 <- GrantNum$TL1
11 KL2 <- GrantNum$KL2
12
13 ## create a single vector with all grant numbers listed
14 AllGrantNums <- as.vector(append(append(UL1, TL1), KL2))
15
16 ## pull out the "kernel" of the grant number, e.g. TR000038
17 kernels <- sapply(AllGrantNums, function(x) substring(x, 5, 13))
18
19 ## create a full grant number without a space, e.g. UL1TR000038
20 fullNums <- gsub(" ", "", AllGrantNums)
21
22 ## create a vector with both the kernel and the full grant to allow
23 ## for searching on different variants of the grant number
24 variants <- as.vector(append(kernels, fullNums))
25
26 publist <- lapply(variants, function(x) entrez_search(db="pubmed", term=x, retmax=5000))
27 pmids <- lapply(publist, function(x) x$sids)
28 pmidlist <- unique(unlist(pmids))
29
30 ## add pmidlist with a pmid with no abstract so that will be an even multiple of 500
31
32
```
- Environment Pane:** Shows the current environment with 178 observations of 6 variables. The variables are GrantNum, UL1, TL1, KL2, and UL1. The values are displayed as a list of character strings.
- Console:** Displays the R version (3.4.1) and the R Foundation for Statistical Computing. It also shows the R license and the RStudio logo. The console output is as follows:

```
R version 3.4.1 (2017-06-30) -- "Single Candle"
Copyright (C) 2017 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.start()' for an HTML browser interface to help,
Type 'q()' to quit R.

[workspace loaded from ~/.RData]

> setwd("~/Google Drive/work/Publication Classification/Analysis/R Code")
> library(rentrez)
>
> GrantNum <- read.csv("~/Google Drive/work/Publication Classification/Analysis/R Code/GrantNums.csv", stringsAsFactors = FALSE, na.strings="")
> UL1 <- GrantNum$UL1
> TL1 <- GrantNum$TL1
> KL2 <- GrantNum$KL2
>
> ## create a single vector with all grant numbers listed
> AllGrantNums <- as.vector(append(append(UL1, TL1), KL2))
>
> |
```
- Navigation Pane:** Shows the R documentation for the 'update.packages' function. The description is as follows:

```
update.packages [utils]

Compare Installed Packages with CRAN-like Repositories

Description
old.packages indicates packages that are installed but not available.
new.packages looks for updates to installed packages.

Usage
update.packages(lib.loc = NULL, repos = getOption("repos"),
  contriburl = contrib.url(repos, type),
  method, instlib = NULL,
  ask = TRUE, available = NULL,
  oldPkgs = NULL, ..., checkBuilt = FALSE,
  type = getOption("pkgType"))

old.packages(lib.loc = NULL, repos = getOption("repos"),
  contriburl = contrib.url(repos, type),
  instPkgs = installed.packages(lib.loc = lib.loc),
  method, available = NULL, checkBuilt = FALSE,
  type = getOption("pkgType"))

new.packages(lib.loc = NULL, repos = getOption("repos"),
  contriburl = contrib.url(repos, type),
  instPkgs = installed.packages(lib.loc = lib.loc),
  method, available = NULL, ask = FALSE, ...)
```

RStudio

- Integrated development environment for R, which means one window with access to:
 - Variables
 - Help window
 - Plots
 - Scripting
- Convenience -- Tab Completion
- Reproducibility -- Integration with R Markdown

The data frame: backbone of R

risk_summary * ProcessDataNotebook.Rmd * EHRData * REDCapData *												
Filter												
	patient_id	age	date_of_test	sex	black	nativeAmerican	asian	pacificIsland	white	height_inches	weight_lbs	temp
1	HHUID0040937	43	2018-01-10	1	0	0	1	0	0	64	164	98
2	HHUID0081981	36	2018-04-01	2	0	0	1	0	0	69	192	97
3	HHUID0030501	33	2017-10-30	1	0	0	1	0	0	63	139	98
4	HHUID0019343	51	2018-04-14	1	1	0	0	0	0	66	181	98
5	HHUID0075313	54	2018-01-21	1	0	0	0	0	1	72	177	98
6	HHUID0083101	45	2017-10-13	1	0	0	0	0	1	67	180	101
7	HHUID0013597	49	2017-11-29	1	1	0	0	0	0	62	146	98
8	HHUID0084797	32	2018-03-25	2	0	0	0	0	1	70	194	97
9	HHUID0023084	38	2018-02-04	2	0	0	0	0	1	75	208	98
10	HHUID0043222	35	2018-02-09	1	0	0	0	0	1	64	151	97
11	HHUID0075792	49	2018-02-06	1	0	0	0	0	1	60	135	98
12	HHUID0085441	51	2018-03-11	1	1	0	0	0	1	69	NA	97
13	HHUID0029155	35	2017-11-26	1	0	0	0	0	1	62	174	98
14	HHUID0029533	52	2018-04-29	1	0	0	0	0	1	69	164	98
15	HHUID0013556	32	2018-04-14	2	0	0	0	0	1	73	216	97
16	HHUID0060753	23	2017-10-18	1	0	0	0	0	1	61	141	98
17	HHUID0040787	39	2018-01-29	1	0	0	0	0	1	70	186	98
18	HHUID0064580	23	2018-05-20	2	1	0	0	0	0	67	209	98
19	HHUID0016042	31	2017-11-14	1	0	0	0	0	1	67	183	98
20	HHUID0067541	55	2018-03-07	2	0	0	1	0	0	73	198	98
21	HHUID0028556	24	2017-09-26	1	0	0	0	0	1	70	186	98
22	HHUID0027191	44	2017-12-30	1	0	0	0	0	1	62	152	97
23	HHUID0010402	42	2017-09-20	2	0	0	0	0	1	67	175	98
24	HHUID0037431	23	2018-01-29	2	0	0	0	0	1	63	167	97
25	HHUID0065432	49	2017-09-23	1	0	0	0	0	1	69	165	98
26	HHUID0019336	36	2017-12-21	1	1	0	0	0	0	70	164	97
27	HHUID0057340	23	2018-04-10	1	0	0	0	0	1	63	155	98

The data frame: backbone of R

Each column is a different **variable**

risk_summary * ProcessDataNotebook.Rmd * EHRData * REDCapData *												
Filter												
	patient_id	age	date_of_test	sex	black	nativeAmerican	asian	pacificIsland	white	height_inches	weight_lbs	temp
1	HHUID0040937	43	2018-01-10	1	0	0	1	0	0	64	164	98
2	HHUID0081981	36	2018-04-01	2	0	0	1	0	0	69	192	98
3	HHUID0030501	33	2017-10-30	1	0	0	1	0	0	63	139	98
4	HHUID0019343	51	2018-04-14	1	1	0	0	0	0	66	181	98
5	HHUID0075313	54	2018-01-21	1	0	0	0	0	1	72	177	98
6	HHUID0083101	45	2017-10-13	1	0	0	0	0	1	67	180	101
7	HHUID0013597	49	2017-11-29	1	1	0	0	0	0	62	146	98
8	HHUID0084797	32	2018-03-25	2	0	0	0	0	1	70	194	98
9	HHUID0023084	38	2018-02-04	2	0	0	0	0	1	75	208	98
10	HHUID0043222	35	2018-02-09	1	0	0	0	0	1	64	151	98
11	HHUID0075792	49	2018-02-06	1	0	0	0	0	1	60	135	98
12	HHUID0085441	51	2018-03-11	1	1	0	0	0	1	69	NA	98
13	HHUID0029155	35	2017-11-26	1	0	0	0	0	1	62	174	98
14	HHUID0029533	52	2018-04-29	1	0	0	0	0	1	69	164	98
15	HHUID0013556	32	2018-04-14	2	0	0	0	0	1	73	216	98
16	HHUID0060753	23	2017-10-18	1	0	0	0	0	1	61	141	98
17	HHUID0040787	39	2018-01-29	1	0	0	0	0	1	70	186	98
18	HHUID0064580	23	2018-05-20	2	1	0	0	0	0	67	209	98
19	HHUID0016042	31	2017-11-14	1	0	0	0	0	1	67	183	98
20	HHUID0067541	55	2018-03-07	2	0	0	1	0	0	73	198	98
21	HHUID0028556	24	2017-09-26	1	0	0	0	0	1	70	186	98
22	HHUID0027191	44	2017-12-30	1	0	0	0	0	1	62	152	98
23	HHUID0010402	42	2017-09-20	2	0	0	0	0	1	67	175	98
24	HHUID0037431	23	2018-01-29	2	0	0	0	0	1	63	167	98
25	HHUID0065432	49	2017-09-23	1	0	0	0	0	1	69	165	98
26	HHUID0019336	36	2017-12-21	1	1	0	0	0	0	70	164	98
27	HHUID0057340	23	2018-04-10	1	0	0	0	0	1	63	155	98

The data frame: backbone of R

Each column is a different **variable**

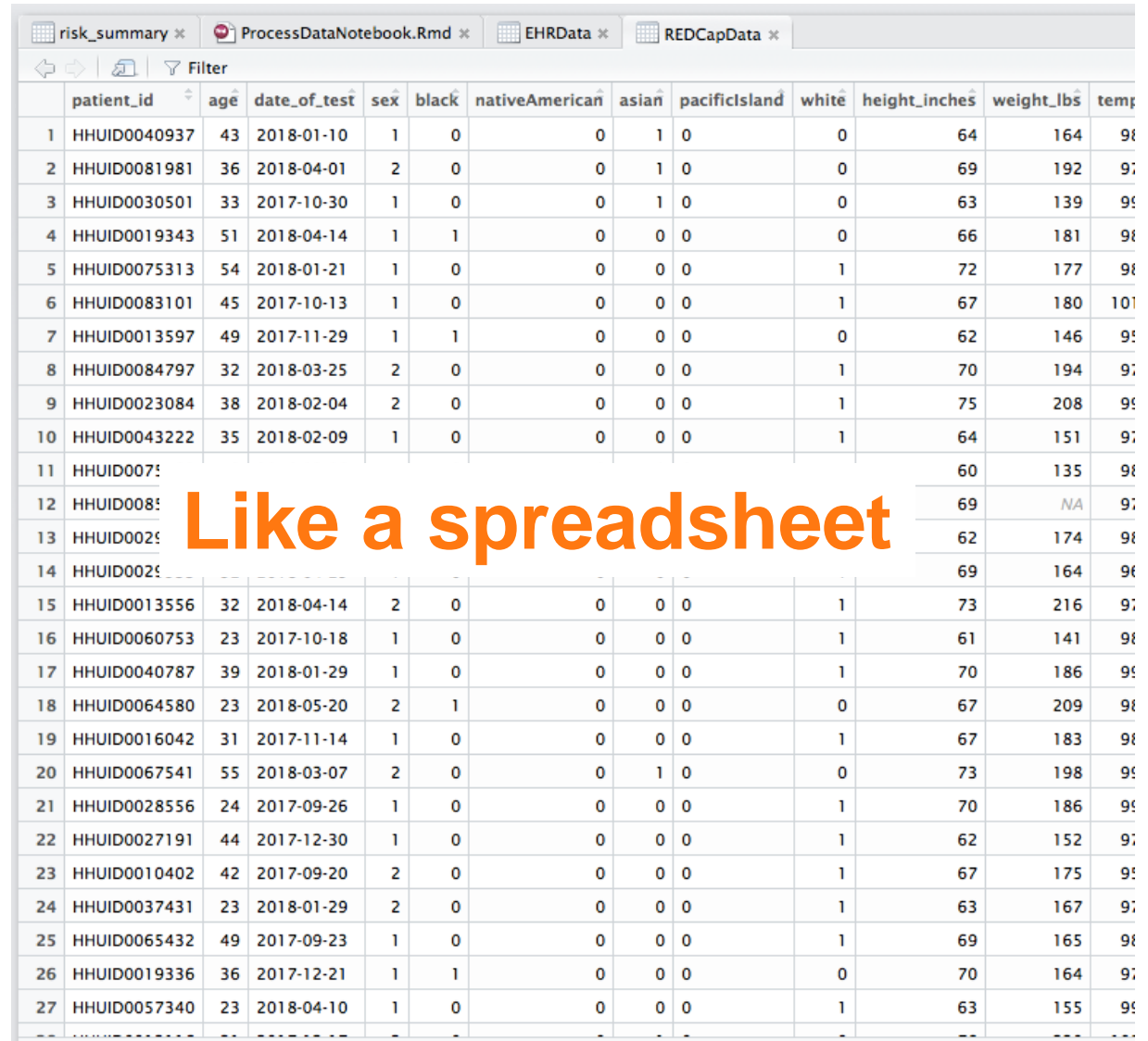
Each row is an **observation**

	patient_id	age	date_of_test	sex	black	nativeAmerican	asian	pacificIsland	white	height_inches	weight_lbs	temp
1	HHUID0040937	43	2018-01-10	1	0	0	1	0	0	64	164	98
2	HHUID0081981	36	2018-04-01	2	0	0	1	0	0	69	192	98
3	HHUID0030501	33	2017-10-30	1	0	0	1	0	0	63	139	98
4	HHUID0019343	51	2018-04-14	1	1	0	0	0	0	66	181	98
5	HHUID0075313	54	2018-01-21	1	0	0	0	0	1	72	177	98
6	HHUID0083101	45	2017-10-13	1	0	0	0	0	1	67	180	101
7	HHUID0013597	49	2017-11-29	1	1	0	0	0	0	62	146	98
8	HHUID0084797	32	2018-03-25	2	0	0	0	0	1	70	194	98
9	HHUID0023084	38	2018-02-04	2	0	0	0	0	1	75	208	98
10	HHUID0043222	35	2018-02-09	1	0	0	0	0	1	64	151	98
11	HHUID0075792	49	2018-02-06	1	0	0	0	0	1	60	135	98
12	HHUID0085441	51	2018-03-11	1	1	0	0	0	1	69	NA	98
13	HHUID0029155	35	2017-11-26	1	0	0	0	0	1	62	174	98
14	HHUID0029533	52	2018-04-29	1	0	0	0	0	1	69	164	98
15	HHUID0013556	32	2018-04-14	2	0	0	0	0	1	73	216	98
16	HHUID0060753	23	2017-10-18	1	0	0	0	0	1	61	141	98
17	HHUID0040787	39	2018-01-29	1	0	0	0	0	1	70	186	98
18	HHUID0064580	23	2018-05-20	2	1	0	0	0	0	67	209	98
19	HHUID0016042	31	2017-11-14	1	0	0	0	0	1	67	183	98
20	HHUID0067541	55	2018-03-07	2	0	0	1	0	0	73	198	98
21	HHUID0028556	24	2017-09-26	1	0	0	0	0	1	70	186	98
22	HHUID0027191	44	2017-12-30	1	0	0	0	0	1	62	152	98
23	HHUID0010402	42	2017-09-20	2	0	0	0	0	1	67	175	98
24	HHUID0037431	23	2018-01-29	2	0	0	0	0	1	63	167	98
25	HHUID0065432	49	2017-09-23	1	0	0	0	0	1	69	165	98
26	HHUID0019336	36	2017-12-21	1	1	0	0	0	0	70	164	98
27	HHUID0057340	23	2018-04-10	1	0	0	0	0	1	63	155	98

The data frame: backbone of R

Each column is a different **variable**

Each row is an
observation



ProcessDataNotebook.Rmd EHRData REDCapData

Filter

	patient_id	age	date_of_test	sex	black	nativeAmerican	asian	pacificIsland	white	height_inches	weight_lbs	temp
1	HHUID0040937	43	2018-01-10	1	0	0	1	0	0	64	164	98
2	HHUID0081981	36	2018-04-01	2	0	0	1	0	0	69	192	98
3	HHUID0030501	33	2017-10-30	1	0	0	1	0	0	63	139	98
4	HHUID0019343	51	2018-04-14	1	1	0	0	0	0	66	181	98
5	HHUID0075313	54	2018-01-21	1	0	0	0	0	1	72	177	98
6	HHUID0083101	45	2017-10-13	1	0	0	0	0	1	67	180	101
7	HHUID0013597	49	2017-11-29	1	1	0	0	0	0	62	146	98
8	HHUID0084797	32	2018-03-25	2	0	0	0	0	1	70	194	98
9	HHUID0023084	38	2018-02-04	2	0	0	0	0	1	75	208	98
10	HHUID0043222	35	2018-02-09	1	0	0	0	0	1	64	151	98
11	HHUID007:									60	135	98
12	HHUID008:									69	NA	98
13	HHUID002:									62	174	98
14	HHUID002:									69	164	98
15	HHUID0013556	32	2018-04-14	2	0	0	0	0	1	73	216	98
16	HHUID0060753	23	2017-10-18	1	0	0	0	0	1	61	141	98
17	HHUID0040787	39	2018-01-29	1	0	0	0	0	1	70	186	98
18	HHUID0064580	23	2018-05-20	2	1	0	0	0	0	67	209	98
19	HHUID0016042	31	2017-11-14	1	0	0	0	0	1	67	183	98
20	HHUID0067541	55	2018-03-07	2	0	0	1	0	0	73	198	98
21	HHUID0028556	24	2017-09-26	1	0	0	0	0	1	70	186	98
22	HHUID0027191	44	2017-12-30	1	0	0	0	0	1	62	152	98
23	HHUID0010402	42	2017-09-20	2	0	0	0	0	1	67	175	98
24	HHUID0037431	23	2018-01-29	2	0	0	0	0	1	63	167	98
25	HHUID0065432	49	2017-09-23	1	0	0	0	0	1	69	165	98
26	HHUID0019336	36	2017-12-21	1	1	0	0	0	0	70	164	98
27	HHUID0057340	23	2018-04-10	1	0	0	0	0	1	63	155	98

Like a spreadsheet

Data types within data frames

- Numeric variable = *numeric* or *integer*

Ex: 1, 1.5, 200000, 3.14159

Data types within data frames

- Numeric variable = *numeric* or *integer*

Ex: 1, 1.5, 200000, 3.14159

- Text variable = *character*

Ex: a, b, hello, 3b

Data types within data frames

- Numeric variable = *numeric* or *integer*

Ex: 1, 1.5, 200000, 3.14159

- Text variable = *character*

Ex: a, b, hello, 3b

- Categorical variable = *factor*

Ex: cat, dog, pig, rhino, horse

Data types within data frames

- Numeric variable = *numeric* or *integer*

Ex: 1, 1.5, 200000, 3.14159

- Text variable = *character*

Ex: a, b, hello, 3b

- Categorical variable = *factor*

Ex: cat, dog, pig, rhino, horse

- Ordered categorical variable = *ordered factor*

Ex: xsmall, small, medium, large, xlarge

Data types within data frames

- Numeric variable = *numeric* or *integer*

Ex: 1, 1.5, 200000, 3.14159

- Text variable = *character*

Ex: a, b, hello, 3b

- Categorical variable = *factor*

Ex: cat, dog, pig, rhino, horse

- Ordered categorical variable = *ordered factor*

Ex: xsmall, small, medium, large, xlarge

- True/false = *logical*

Ex: TRUE, FALSE

Talking to R

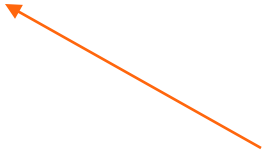
Function calls

- how to tell R what to do
- always followed by ()

`getwd()`

Functions: Output


```
dat <- read.csv(file = "allFound.csv", header = TRUE)
```



Assignment operator (<-): assigns the output of the function to a name that you can then refer back to

Functions: Input

```
dat <- read.csv(file = "allFound.csv", header = TRUE)
```



Arguments: instructions that specify how a function should be run. Not always required, may be more than one, they are separated by commas, and they all go in the ()

Functions: Input

```
dat <- read.csv(file = "allFound.csv", header = TRUE)
```



required



optional

Argument: instructions that specify how a function should be run. Not always required, may be more than one, they are separated by commas, and they all go in the ()

Functions: Output

```
dat <- read.csv(file = "allFound.csv", header = TRUE)
```

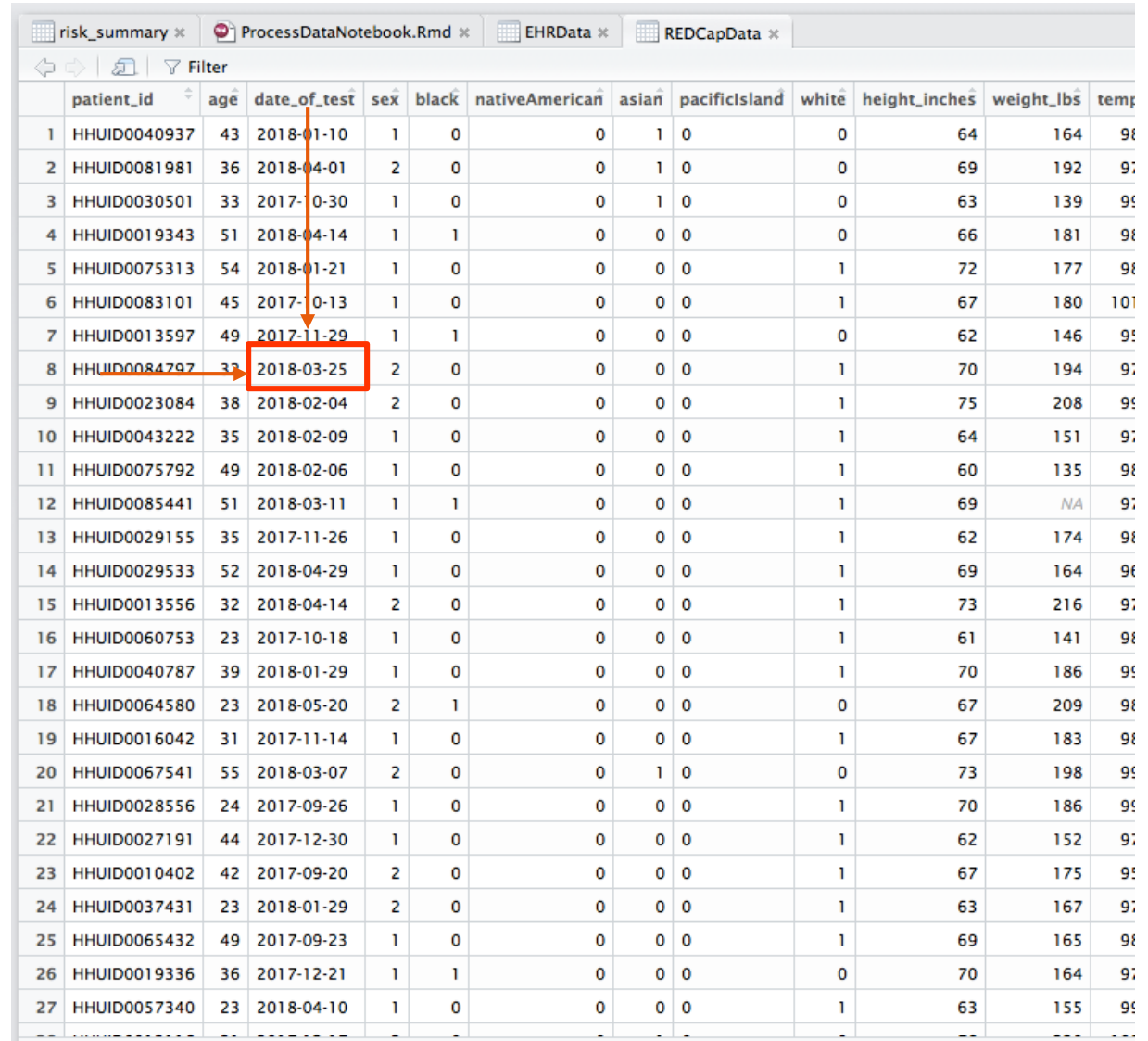


Data frame

Data frames: Accessing elements

Each column is a different **variable**

Each row is an **observation**



	patient_id	age	date_of_test	sex	black	nativeAmerican	asian	pacificIsland	white	height_inches	weight_lbs	temp
1	HHUID0040937	43	2018-01-10	1	0	0	1	0	0	64	164	98
2	HHUID0081981	36	2018-04-01	2	0	0	1	0	0	69	192	98
3	HHUID0030501	33	2017-0-30	1	0	0	1	0	0	63	139	98
4	HHUID0019343	51	2018-04-14	1	1	0	0	0	0	66	181	98
5	HHUID0075313	54	2018-01-21	1	0	0	0	0	1	72	177	98
6	HHUID0083101	45	2017-0-13	1	0	0	0	0	1	67	180	101
7	HHUID0013597	49	2017-11-29	1	1	0	0	0	0	62	146	98
8	HHUID0084297	37	2018-03-25	2	0	0	0	0	1	70	194	98
9	HHUID0023084	38	2018-02-04	2	0	0	0	0	1	75	208	98
10	HHUID0043222	35	2018-02-09	1	0	0	0	0	1	64	151	98
11	HHUID0075792	49	2018-02-06	1	0	0	0	0	1	60	135	98
12	HHUID0085441	51	2018-03-11	1	1	0	0	0	1	69	NA	98
13	HHUID0029155	35	2017-11-26	1	0	0	0	0	1	62	174	98
14	HHUID0029533	52	2018-04-29	1	0	0	0	0	1	69	164	98
15	HHUID0013556	32	2018-04-14	2	0	0	0	0	1	73	216	98
16	HHUID0060753	23	2017-10-18	1	0	0	0	0	1	61	141	98
17	HHUID0040787	39	2018-01-29	1	0	0	0	0	1	70	186	98
18	HHUID0064580	23	2018-05-20	2	1	0	0	0	0	67	209	98
19	HHUID0016042	31	2017-11-14	1	0	0	0	0	1	67	183	98
20	HHUID0067541	55	2018-03-07	2	0	0	1	0	0	73	198	98
21	HHUID0028556	24	2017-09-26	1	0	0	0	0	1	70	186	98
22	HHUID0027191	44	2017-12-30	1	0	0	0	0	1	62	152	98
23	HHUID0010402	42	2017-09-20	2	0	0	0	0	1	67	175	98
24	HHUID0037431	23	2018-01-29	2	0	0	0	0	1	63	167	98
25	HHUID0065432	49	2017-09-23	1	0	0	0	0	1	69	165	98
26	HHUID0019336	36	2017-12-21	1	1	0	0	0	0	70	164	98
27	HHUID0057340	23	2018-04-10	1	0	0	0	0	1	63	155	98

Data frames: Accessing elements

Each column is a different **variable**

Each row is an
observation

`dat[rows, columns]`

Data frames: Accessing elements

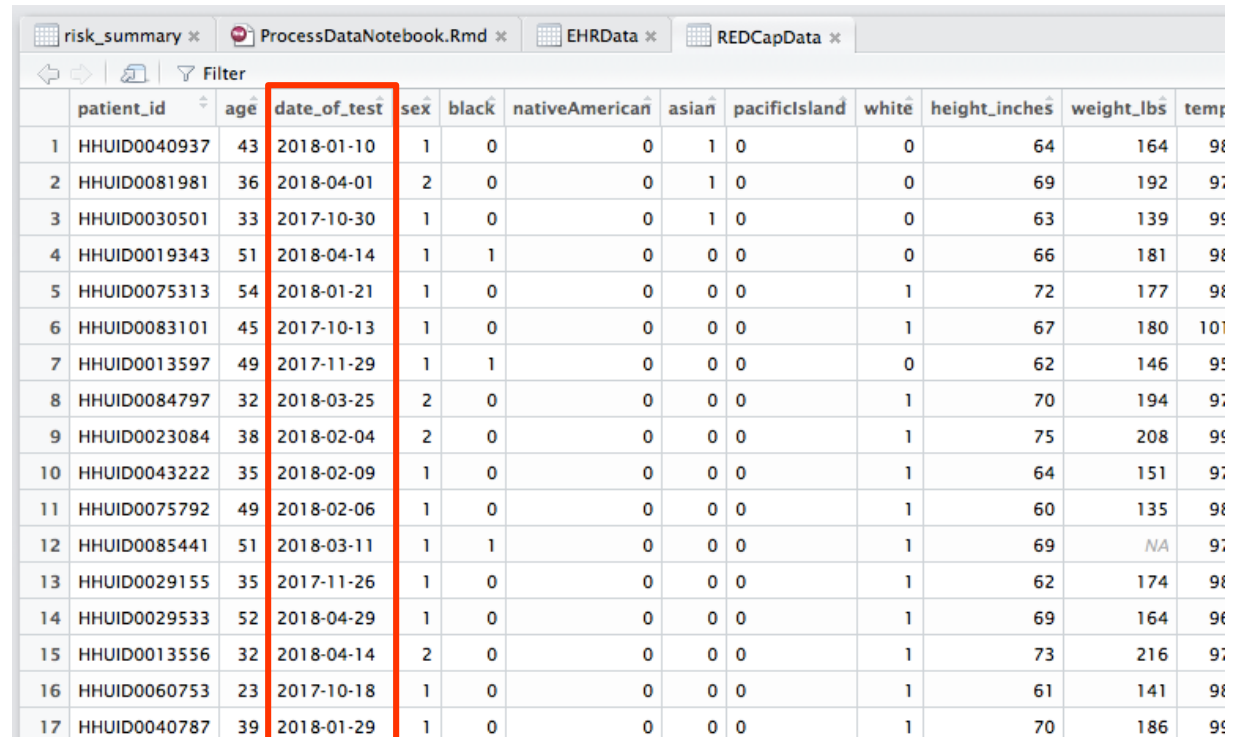
Each column is a different **variable**

Each row is an
observation

`dat[rows, columns]`

`dat[observations, variables]`

Data Frames: Accessing variables



The screenshot shows a data frame with 17 rows and 13 columns. The columns are: patient_id, age, date_of_test, sex, black, nativeAmerican, asian, pacificIsland, white, height_inches, weight_lbs, and temp. The 'date_of_test' column is highlighted with a red box. The data is as follows:

	patient_id	age	date_of_test	sex	black	nativeAmerican	asian	pacificIsland	white	height_inches	weight_lbs	temp
1	HHUID0040937	43	2018-01-10	1	0	0	1	0	0	64	164	98
2	HHUID0081981	36	2018-04-01	2	0	0	1	0	0	69	192	98
3	HHUID0030501	33	2017-10-30	1	0	0	1	0	0	63	139	98
4	HHUID0019343	51	2018-04-14	1	1	0	0	0	0	66	181	98
5	HHUID0075313	54	2018-01-21	1	0	0	0	0	1	72	177	98
6	HHUID0083101	45	2017-10-13	1	0	0	0	0	1	67	180	101
7	HHUID0013597	49	2017-11-29	1	1	0	0	0	0	62	146	98
8	HHUID0084797	32	2018-03-25	2	0	0	0	0	1	70	194	98
9	HHUID0023084	38	2018-02-04	2	0	0	0	0	1	75	208	98
10	HHUID0043222	35	2018-02-09	1	0	0	0	0	1	64	151	98
11	HHUID0075792	49	2018-02-06	1	0	0	0	0	1	60	135	98
12	HHUID0085441	51	2018-03-11	1	1	0	0	0	1	69	NA	98
13	HHUID0029155	35	2017-11-26	1	0	0	0	0	1	62	174	98
14	HHUID0029533	52	2018-04-29	1	0	0	0	0	1	69	164	98
15	HHUID0013556	32	2018-04-14	2	0	0	0	0	1	73	216	98
16	HHUID0060753	23	2017-10-18	1	0	0	0	0	1	61	141	98
17	HHUID0040787	39	2018-01-29	1	0	0	0	0	1	70	186	98

`dat$date_of_test`

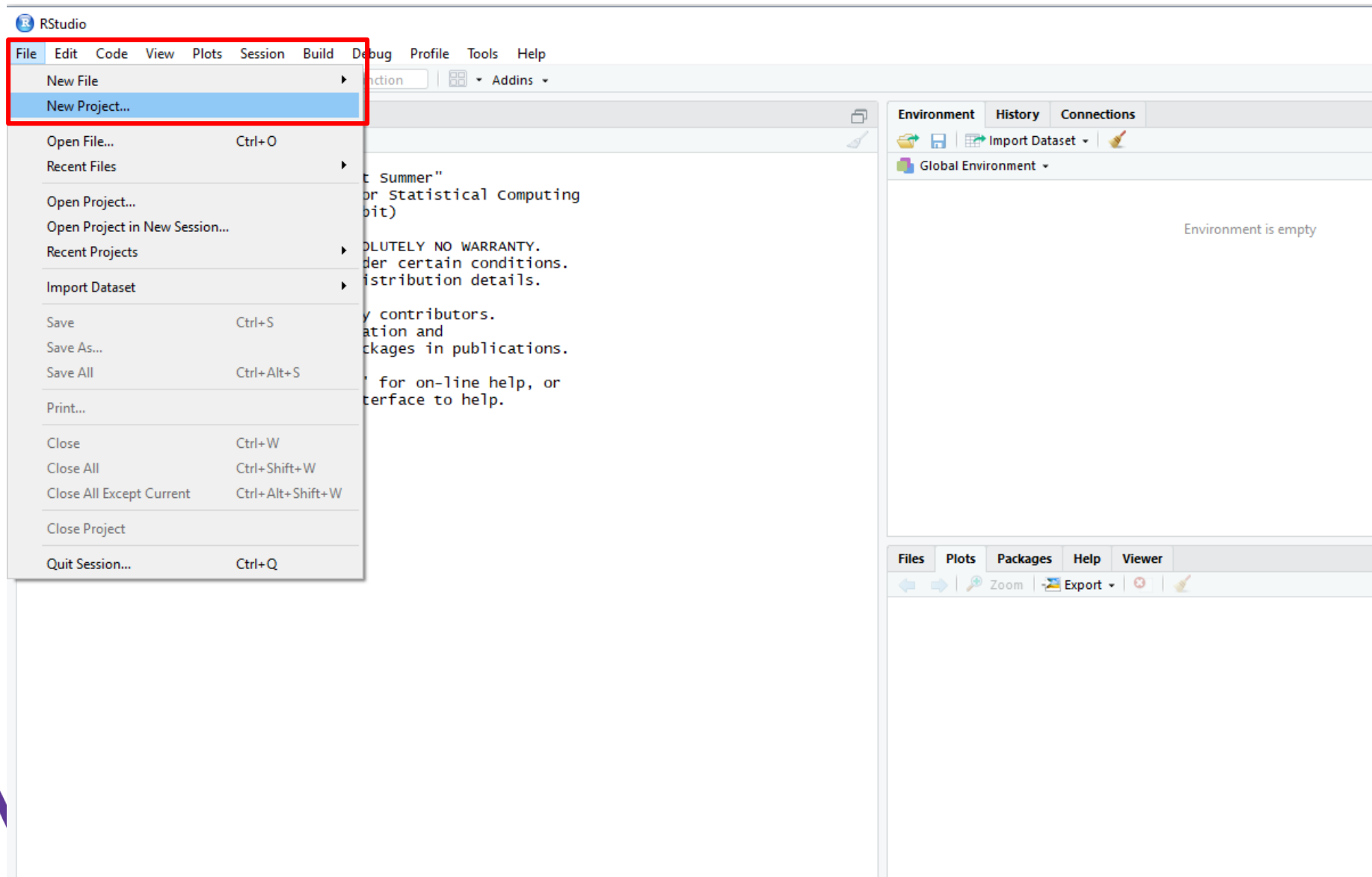
R Projects

- An R Project has its own
 - Working directory
 - Workspace
 - History
 - Source documents
- Convenience of starting where you left off
- Helps with organization

Creating your R Project

- Open R Studio
- Select New Project from the File menu

Creating an R Project



Creating an R Project

The screenshot displays the RStudio interface. 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 opening files, saving, and navigating. The main console window shows the R version 3.4.2 (2017-09-28) -- "short Summer" and copyright information. The 'New Project' dialog box is open, showing three options: 'New Directory' (highlighted with a red box), 'Existing Directory', and 'Version Control'. The 'New Directory' option is described as 'Start a project in a brand new working directory'. The 'Existing Directory' option is described as 'Associate a project with an existing working directory'. The 'Version Control' option is described as 'Checkout a project from a version control repository'. The 'Cancel' button is at the bottom right of the dialog box. The right-hand pane shows the 'Environment' tab with 'Global Environment' and 'Environment is empty'.

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function Addins

Console Terminal

c:/RStudio/

R version 3.4.2 (2017-09-28) -- "short Summer"
Copyright (C) 2017 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. Type 'contributors()' for more details.
'citation()' on how to cite R in publications.


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


> |

New Project

Create Project

 **New Directory** >
Start a project in a brand new working directory

 **Existing Directory** >
Associate a project with an existing working directory

 **Version Control** >
Checkout a project from a version control repository

Cancel

Environment History Connections

Import Dataset

Global Environment

Environment is empty

Files Plots Packages Help Viewer

Zoom Export

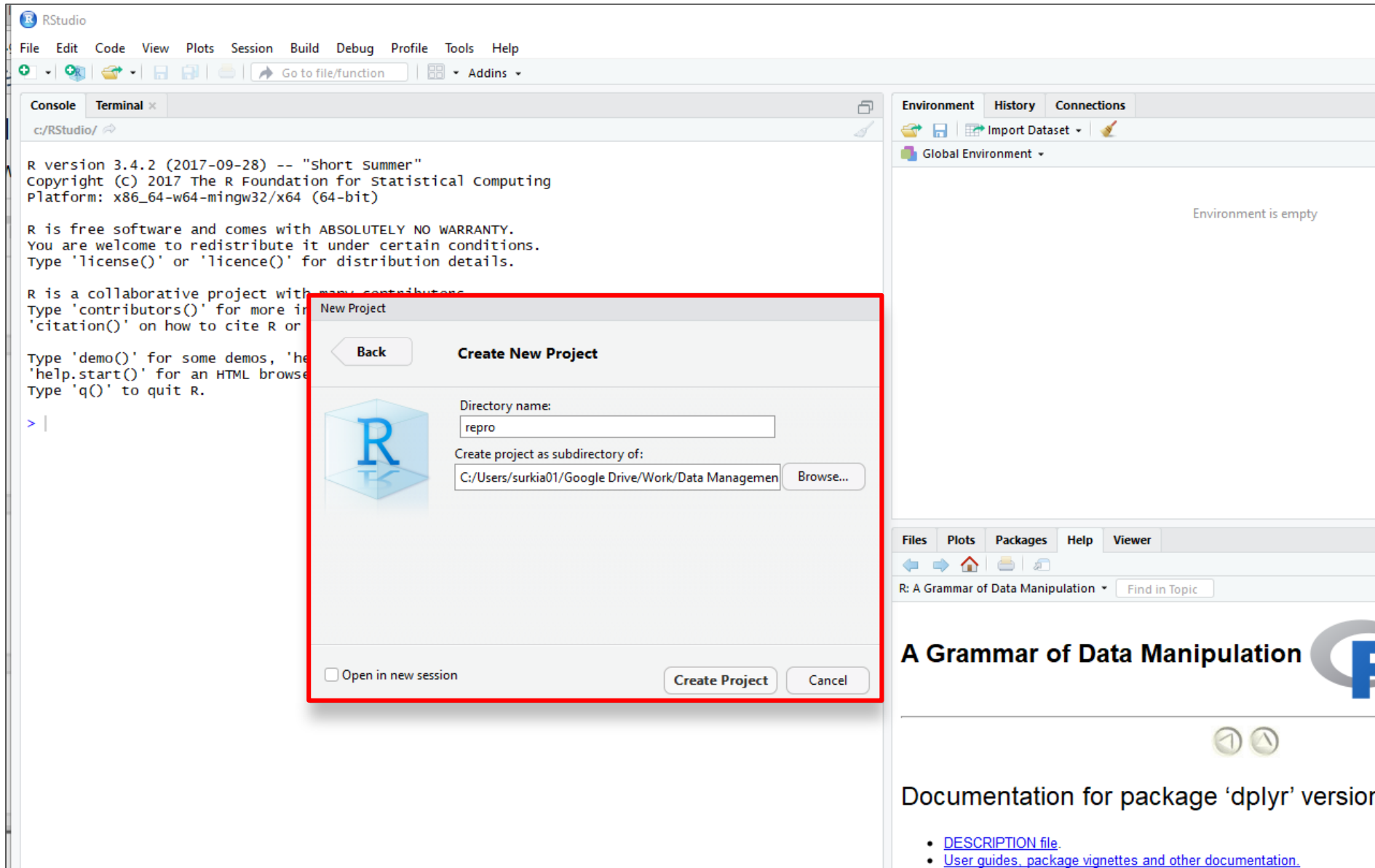
Creating an R Project

The screenshot shows the RStudio interface with the 'New Project' dialog box open. The dialog box has a 'Back' button and a 'Project Type' section. The 'Project Type' section lists several options, with 'New Project' highlighted by a red rectangle. The options are:

- New Project
- R Package
- Shiny Web Application
- R Package using Rcpp
- R Package using RcppArmadillo
- R Package using RcppEigen
- R Package using devtools

The 'Console' pane on the left shows the R version 3.4.2 (2017-09-28) -- "short Summer" and the copyright information. The 'Environment' pane on the right shows 'Global Environment' and 'Environment is empty'. The 'Files' pane at the bottom shows 'Files', 'Plots', 'Packages', 'Help', and 'Viewer' tabs.

Creating an R Project



The screenshot displays the RStudio interface with the 'New Project' dialog box open. The dialog box is titled 'New Project' and has a 'Back' button. It contains the following fields and options:

- Directory name:** A text box containing the text 'repro'.
- Create project as subdirectory of:** A text box containing the path 'C:/Users/surkia01/Google Drive/Work/Data Managemen', followed by a 'Browse...' button.
- Open in new session:** An unchecked checkbox.
- Buttons:** 'Create Project' and 'Cancel' buttons at the bottom right.

The background shows the RStudio console with the R version 3.4.2 (2017-09-28) -- "Short Summer" message, and the Environment pane on the right showing 'Global Environment'.

A Grammar of Data Manipulation








Documentation for package 'dplyr' version

- [DESCRIPTION file.](#)
- [User guides, package vignettes and other documentation.](#)

Structuring your project

Create the following folders within your repro folder:

- code
- raw_data
- processed_data
- results

Name	Date modified	Type	Size
 .Rproj.user	7/2/2018 3:03 PM	File folder	
 code	7/2/2018 3:04 PM	File folder	
 processed_data	7/2/2018 3:04 PM	File folder	
 raw_data	7/2/2018 3:04 PM	File folder	
 results	7/2/2018 3:04 PM	File folder	
 desktop.ini	7/2/2018 3:03 PM	Configuration sett...	1 KB
 repro.Rproj	7/2/2018 3:03 PM	RPROJ File	1 KB

Data files

Drag and drop the following two files into the code folder that is inside your repro folder:

- EHRData.csv
- REDCapData_Export.csv

Good Enough Practices for Scientific Computing: Data Management

- Save the raw data
- Record all steps for processing data
 - Do all processing within R
- Each project in its own directory
 - Directory named after project
- Each document type in its own sub-directory
- All files with names that reflect their content/function

GO TO HANDOUT

Reproducibility Checklist

Documentation

- ☐ Is it clear where to begin? (e.g., can someone picking a project up see where to start running it)
- ☐ Can you determine which file(s) was/were used as input in a process that produced a derived file?
- ☐ Is there documentation about every result?
- ☐ Have you noted the exact version of every external application used in the process?
- ☐ For analyses that include randomness, have you noted the underlying random seed(s)?

Reproducibility Checklist

Organization

- ☐ Which is the most recent data file/code?
- ☐ Which folders can I safely delete?
- ☐ Have you stored the raw data behind each plot?
- ☐ Do you run backups on all files associated with your analysis?

Reproducibility Checklist

Automation

- ☐ Are there lots of manual data manipulation steps are there?
- ☐ Are all custom scripts under version control?
- ☐ Is your writing (content) under version control?

Reproducibility Checklist

Publication

- ☐ Have you archived the exact version of every external application used in your process(es)?
- ☐ Are textual statements connected/linked to the supporting results or data?
- ☐ Did you archived preprints of resulting papers in a public repository?
- ☐ Did you release the underlying code at the time of publishing a paper?
- ☐ Are you providing public access to your scripts, runs, and results?