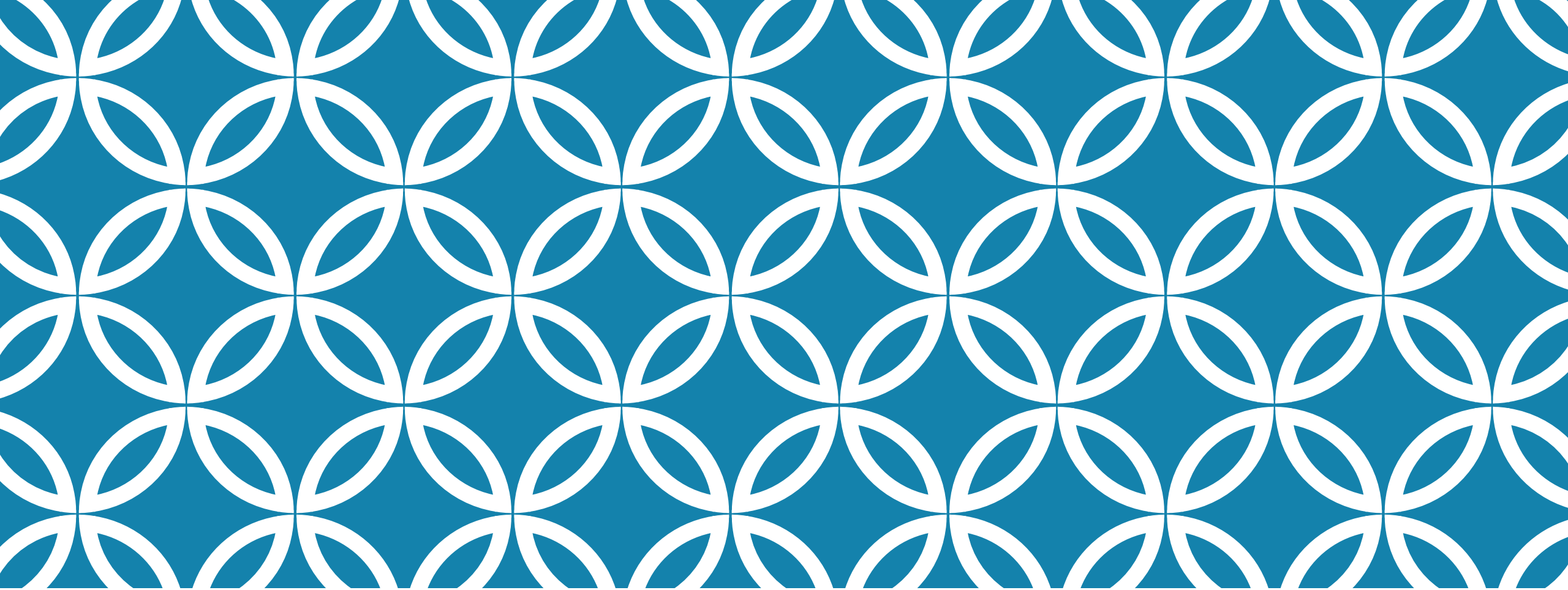


Reproducible Clinical Data Analysis with R and RStudio

Session 4

Reproducible Reports

April 5, 2018

Introduction	Early: 3/27, 4 PM Late: 3/27, 5:30 PM
Getting Data	Early: 3/28, 4 PM Late: 3/28, 5:30 PM
Exploring Data	Early: 4/3, 4 PM Late: 4/2, 5:30 PM
Reproducible Reports	Early: 4/5, 4 PM Late: 4/5, 5:30 PM
Course Project Presentations	TBA

R Markdown and R Notebooks

R Markdown Reference Guide

Learn more about R Markdown at rmarkdown.rstudio.com
Learn more about Interactive Docs at shiny.rstudio.com/articles

Contents:

Syntax

```
Plain text

End a line with two spaces
to start a new paragraph.

*italics* and _italics_

**bold** and __bold__

superscript^2^

~~strikethrough~~

[link](www.rstudio.com)

# Header 1

## Header 2

### Header 3

#### Header 4
```

Becomes

Plain text

End a line with two spaces to start a new paragraph.

italics and *italics*

bold and **bold**

superscript^2

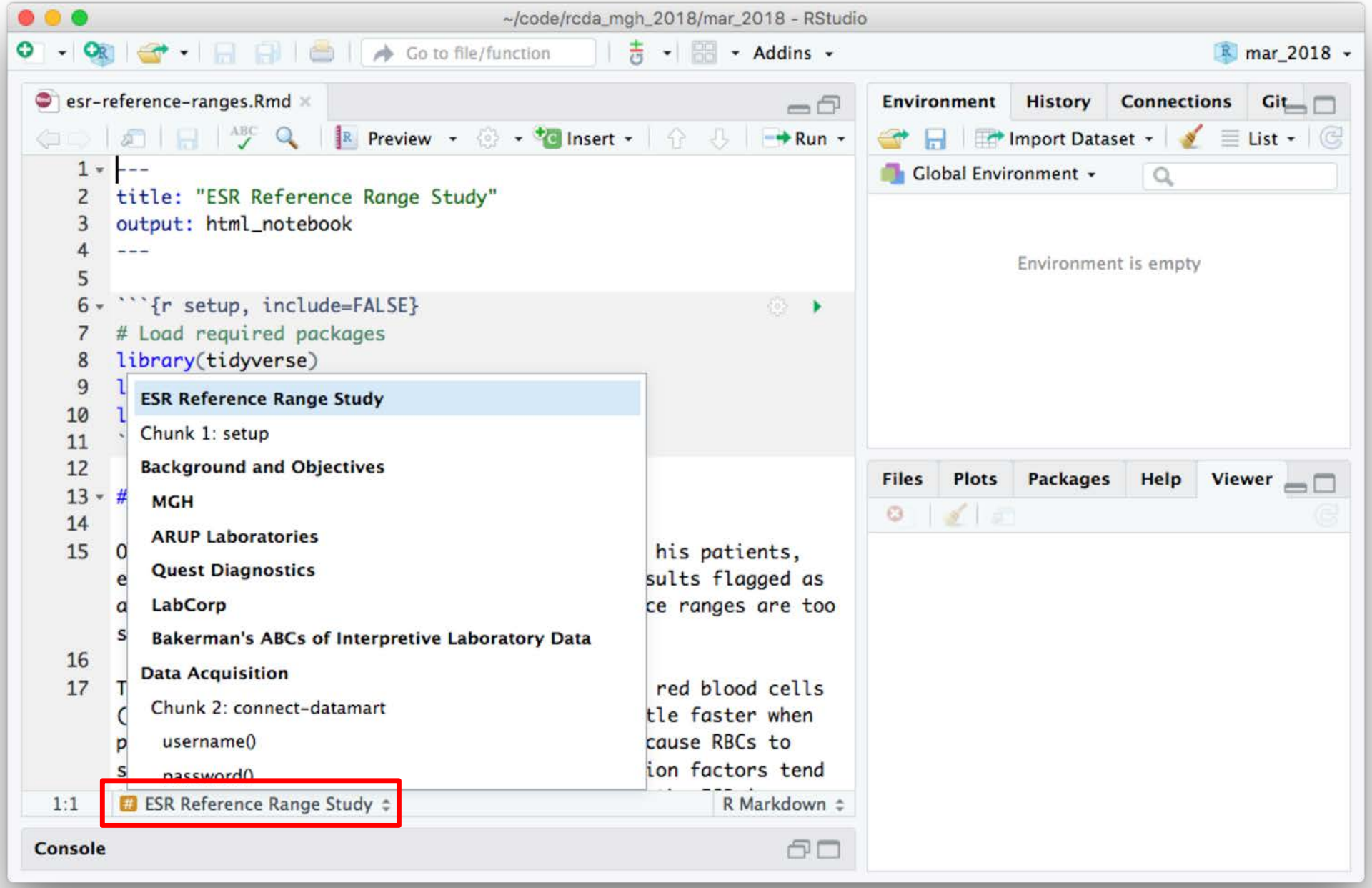
~~strikethrough~~

link

Header 1

Header 2

Header 3


~/code/rcda_mgh_2018/mar_2018 - RStudio
Go to file/function
Addins
mar_2018
esr-reference-ranges.Rmd
Preview
Insert
Run
1 ---
2 title: "ESR Reference Range Study"
3 output: html_notebook
4 ---
5
6 ```{r setup, include=FALSE}
7 # Load required packages
8 library(tidyverse)
ESR Reference Range Study
Chunk 1: setup
Background and Objectives
MGH
ARUP Laboratories
Quest Diagnostics
LabCorp
Bakerman's ABCs of Interpretive Laboratory Data
Data Acquisition
Chunk 2: connect-datamart
username()
password()
his patients,
sults flagged as
ce ranges are too
red blood cells
tle faster when
cause RBCs to
ion factors tend
1:1
ESR Reference Range Study
R Markdown
Console
Environment
History
Connections
Git
Import Dataset
List
Global Environment
Environment is empty
Files
Plots
Packages
Help
Viewer

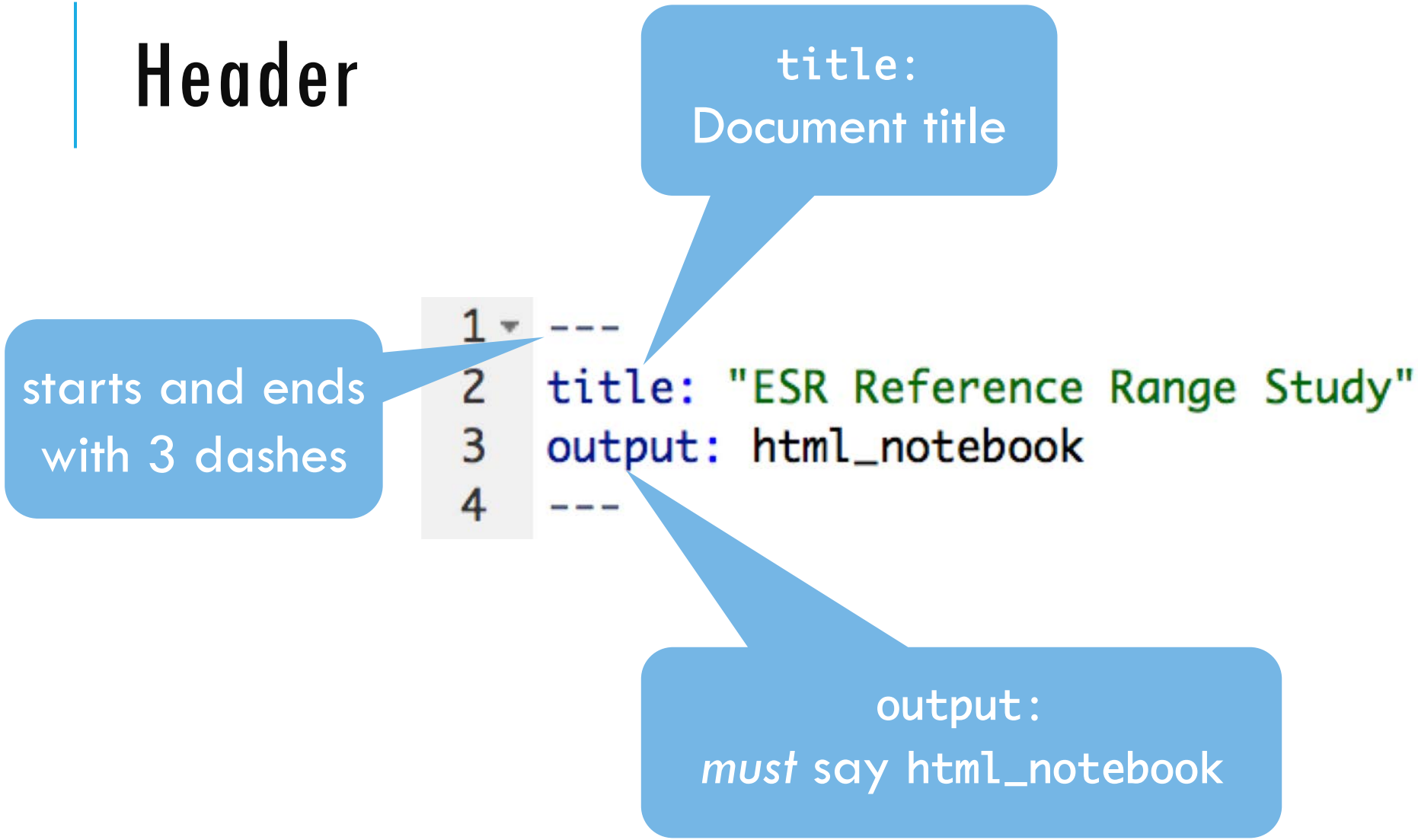
Your Turn #1

Create a new R Notebook. Save it on your Desktop as `Practice.Rmd` and preview it.

Now delete all the contents below the header and try to write the R Markdown that recapitulates the handout titled "R Notebook Practice".

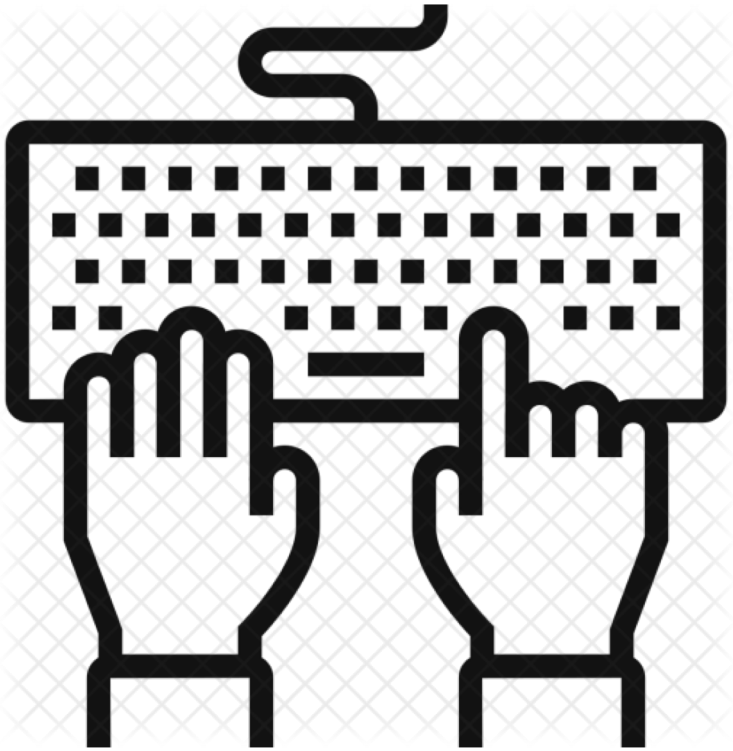
Anatomy of a Reproducible Report

Header


title:
Document title
starts and ends with 3 dashes
---
title: "ESR Reference Range Study"
output: html_notebook
---
output:
must say html_notebook

“Copy/Paste” Code

“Write Yourself” Code



"Setup" Chunk

chunk name (optional)

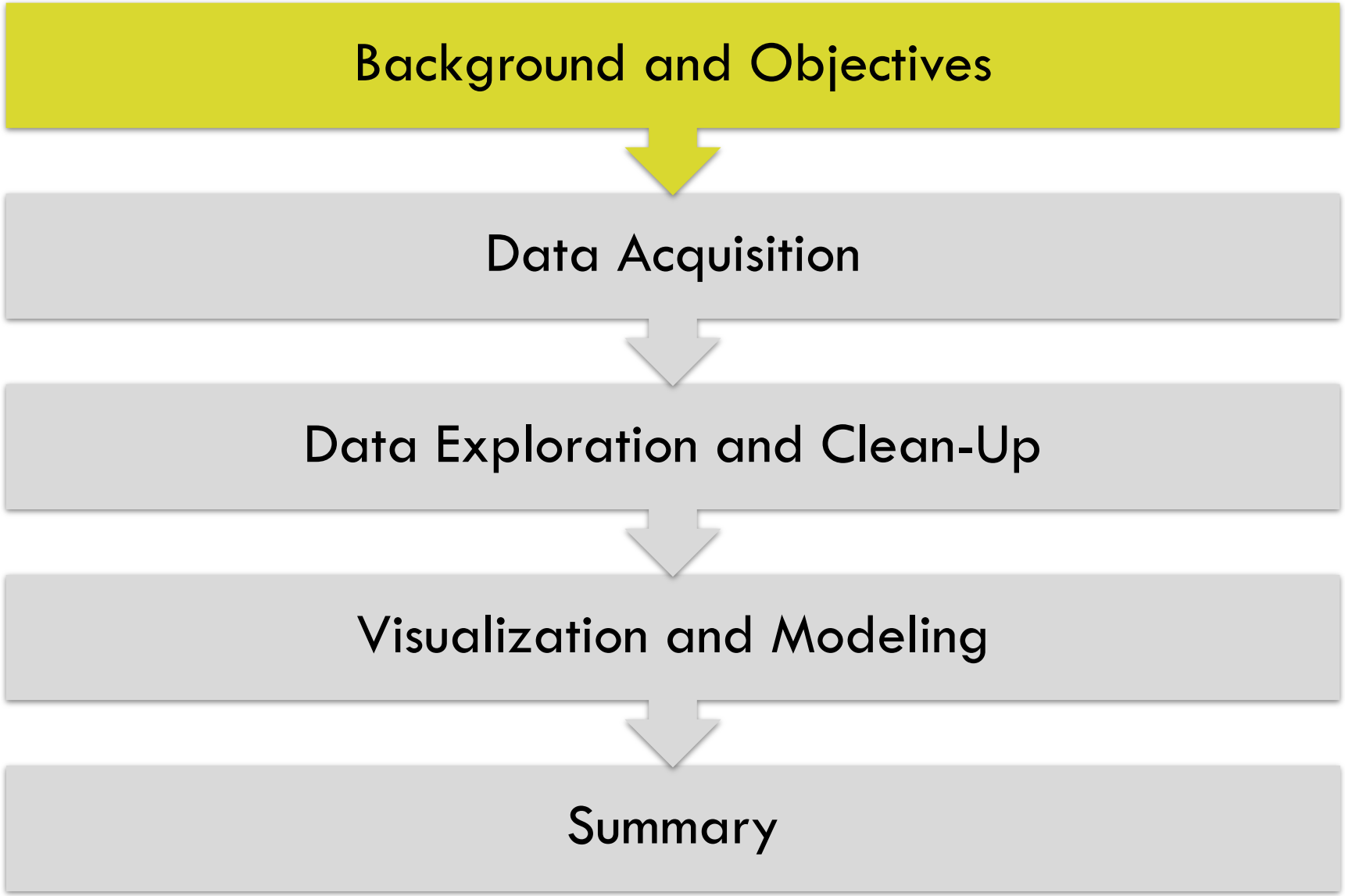
chunk option: don't show code in rendered document

comments start with #

````
```{r setup, include=FALSE}
# Load required packages
library(tidyverse)
library(odbc)
library(lubridate)
```
````

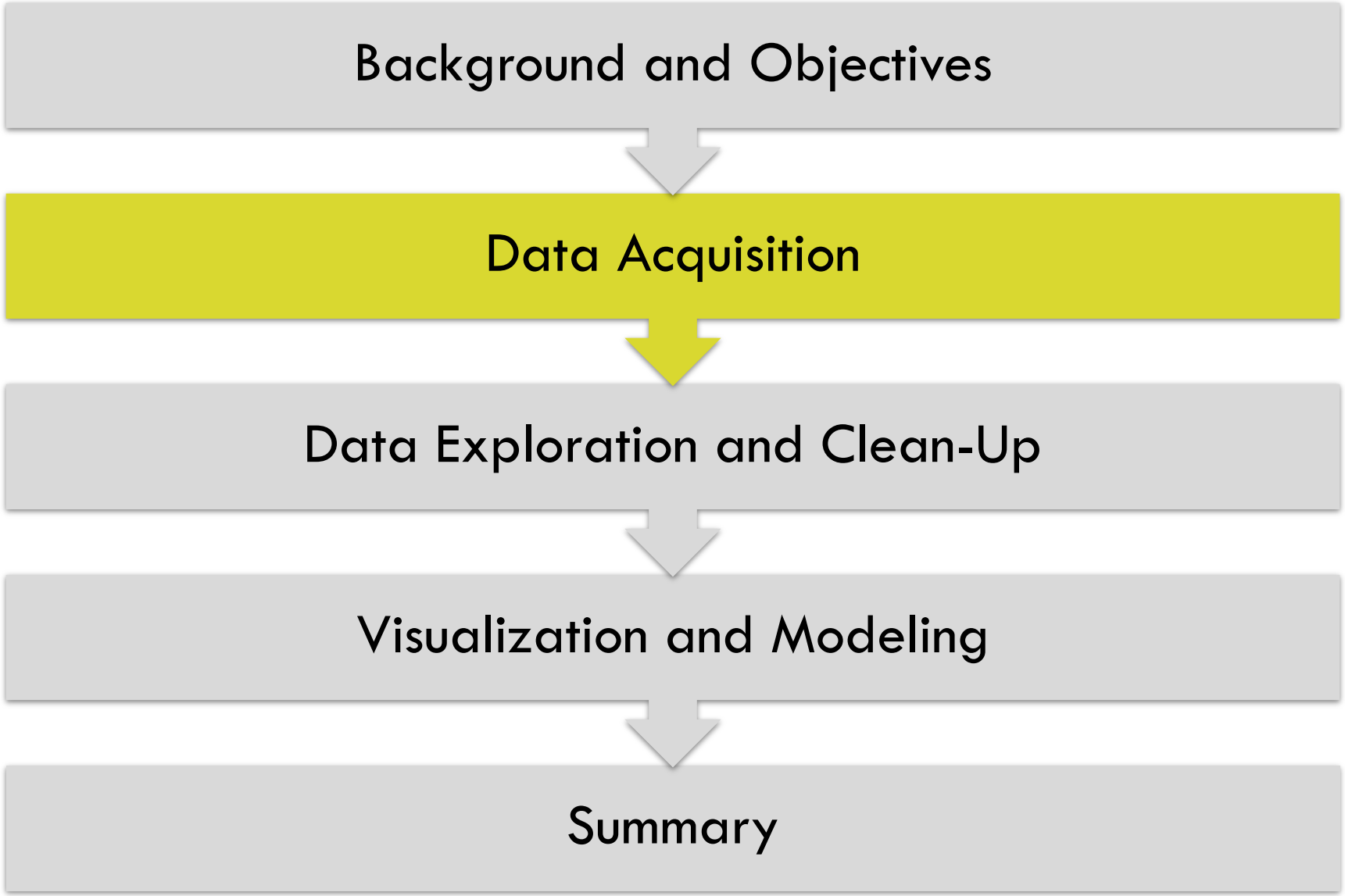
for database access

for dealing with dates


Background and Objectives
Data Acquisition
Data Exploration and Clean-Up
Visualization and Modeling
Summary

Background and Objectives

- What is the question to be addressed by the project?
- How could addressing this question change patient management or further generalizable knowledge?
- What data resources could be used?
- What are the *first 1-2 key graphs or statistics* that should be obtained?


Background and Objectives
Data Acquisition
Data Exploration and Clean-Up
Visualization and Modeling
Summary

Data Acquisition

- Narrative
  - Database sources
    - Database server address and database table name
  - Data file sources
    - Origin, with references
    - File name and format
  - Both
    - General properties of the data set
    - Code book for variables
- Code
  - Database sources
    - Connection and query
      - Don't hard-code credentials!
    - Data dump
    - Disconnect
  - Data file sources
    - Code to read data file

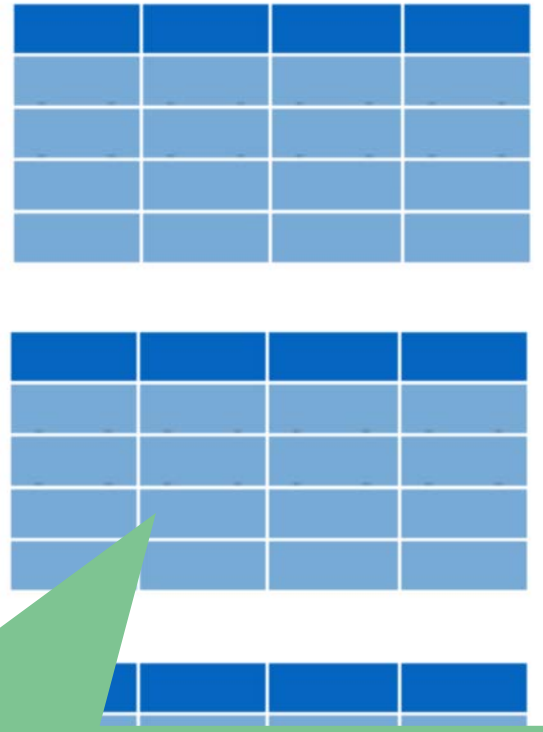
Database Management System

PHSSQL2057
("Datamart")

Database

MGHLABUTIL

Database Tables



MGHLABUTIL_LabResults

Connect to Datamart

functions to ask login credentials

connection handle

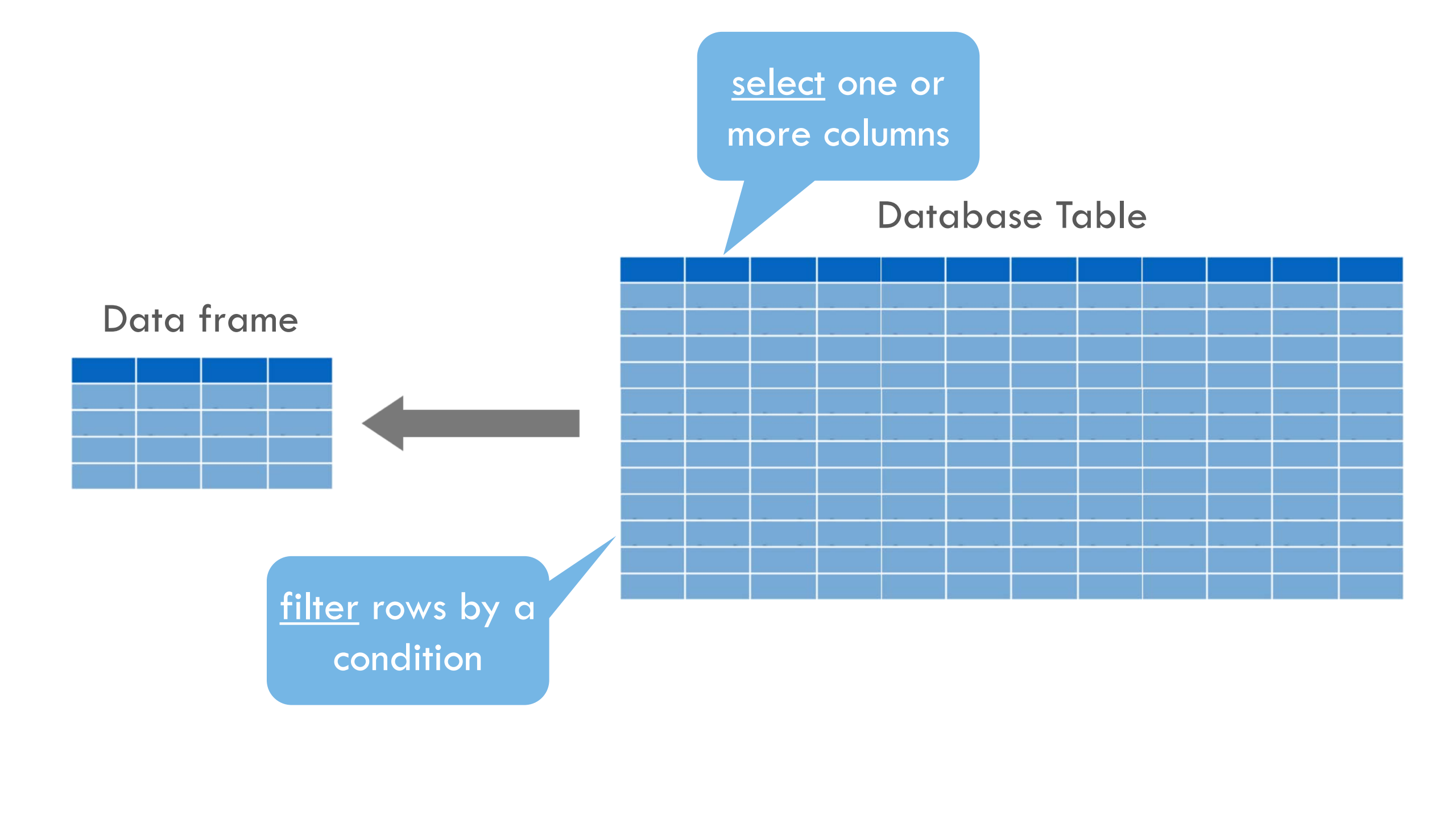
pointer to database table

```r
```{r connect-datamart}
username <- function() {
    .rs.api.showPrompt(title = "Username",
                       message = "Please enter your Partners User Name")
}

password <- function() {
    .rs.api.askForPassword(prompt = "Please enter your Partners Password")
}

con <- dbConnect(odbc(),
                 driver = "{FreeTDS}",
                 dsn = "PHSSQL2057",
                 uid = str_c("PARTNERS\\", username()),
                 pwd = password())

labresults_tbl <- tbl(con, "MGHLABUTIL_LabResults")
```
```


select one or more columns
Database Table
Data frame
filter rows by a condition

Your Turn #2

Connect to the Datamart by running the code chunk labeled `connect-datamart`.

Find the `MGHLABUTIL` database and within it, find the `MGHLABUTIL_LabResults` table.

Preview it by clicking the symbol that looks like a small table.

There are 71 columns in this database table. Pair up and discuss which specific columns you might want to pull to examine ESR reference ranges.

Code Book

Retrieve all ESR values from 2017, along with the following columns:

1. **CollAge**. Patient age at time of collection.
2. **CollectDateTime**. Date and time of collection.
3. **PtName**. Patient name.
4. **PtNumber**. Patient MRN (MGH).
5. **PtSex**. Patient sex.
6. **Result**. ESR result value.
7. **TstOrderName**. Test order name.
8. **UserQAFlags**. This column marks “high” and “low” results.

Database Query

You can treat a database table pointer like a data frame!

```
96  ```{r query}
97  esr <- labresults_tbl %>%
98    select(CollAge,
99           CollectDateTime,
100          PtName,
101          PtNumber,
102          PtSex,
103          Result,
104          TstOrderName,
105          UserQAFlags) %>%
106   filter(TstOrderName == "ESR",
107          year(CollectDateTime) == "2017") %>%
108   arrange(CollectDateTime) %>%
109   collect()
110 ```
```

collect():
pull all results
(By default, max. 1000)

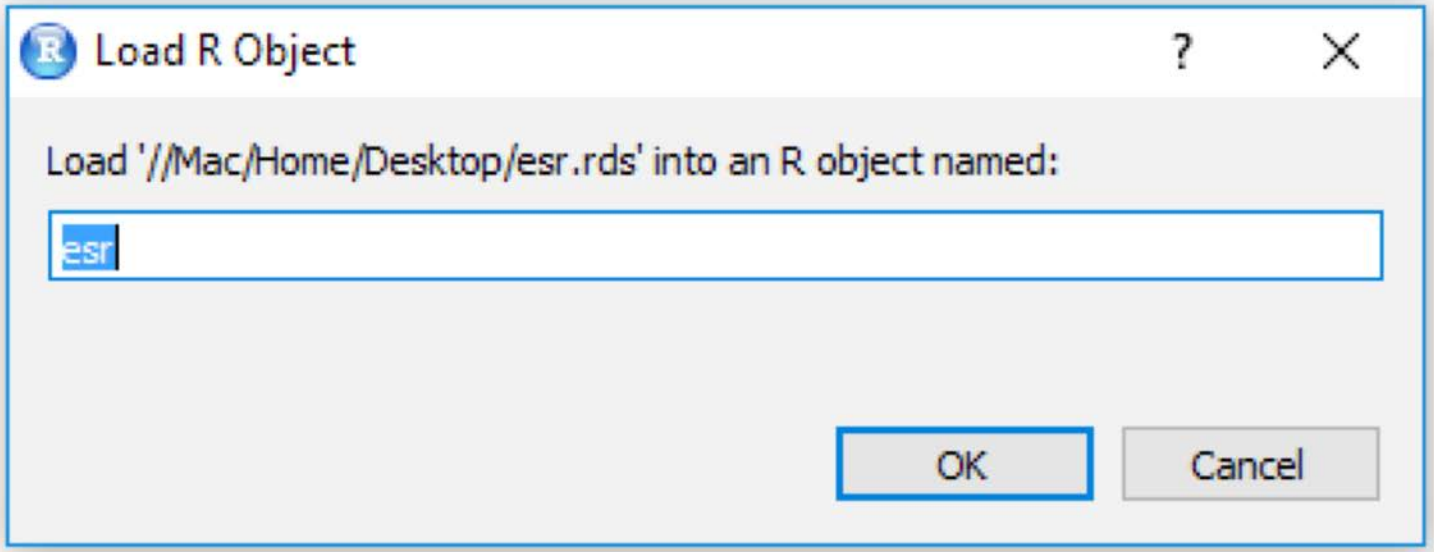
Save Data Dump

str_c(...):
join multiple strings
into one string

````
110 ```{r save-rds}
111 esr %>%
112     write_rds(str_c("esr_data_", today(), ".rds"))
113 ```
````

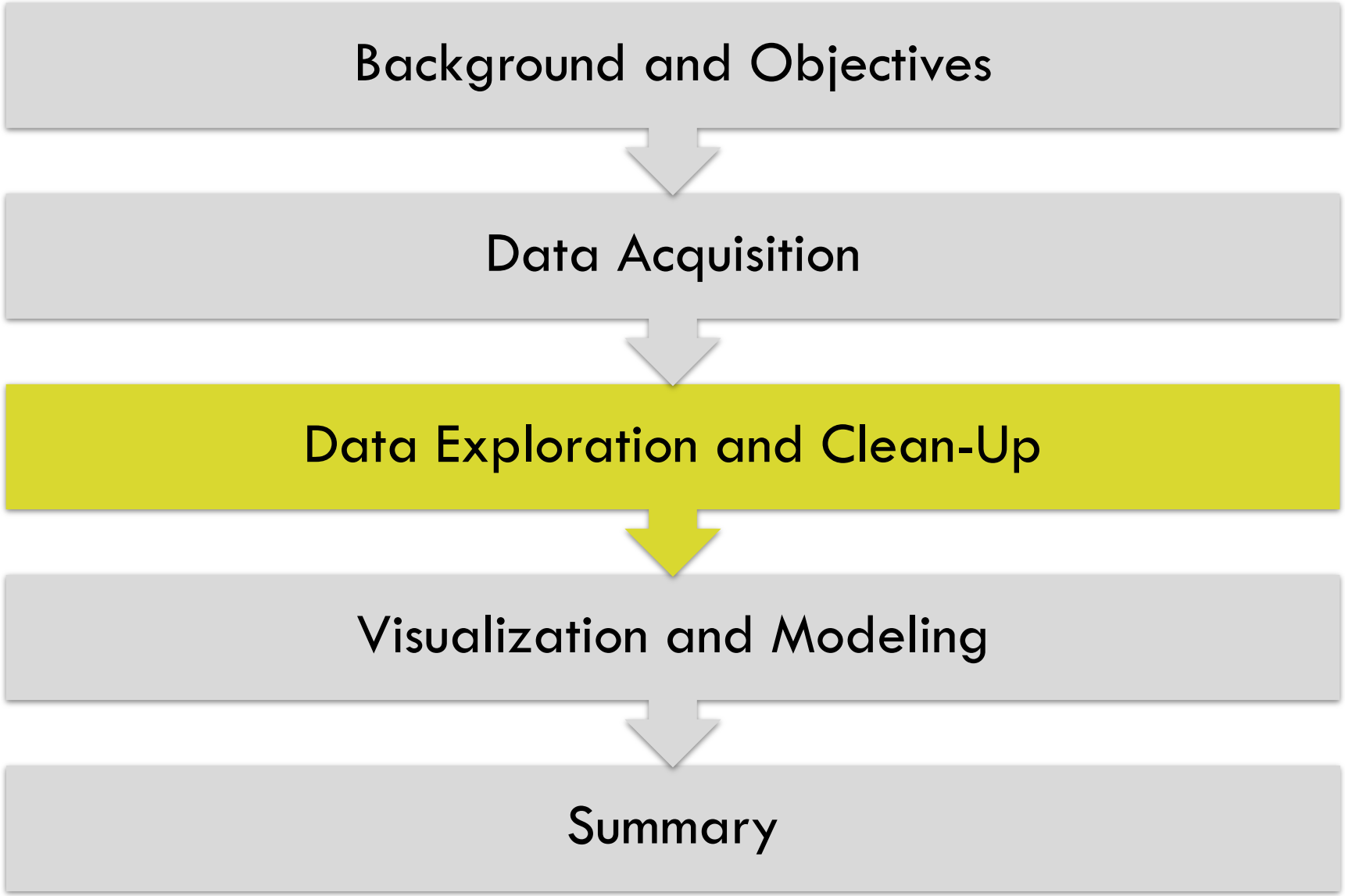
esr_data_2018-04-05.rds

Load Data Dump (if necessary)



Disconnect from Database

````
117 ```{r}
118 dbDisconnect(con)
119 ```
````


Background and Objectives
Data Acquisition
Data Exploration and Clean-Up
Visualization and Modeling
Summary

Data Exploration and Clean-Up

- Convert messy data to tidy
- Check for missing values
- Verify (and if necessary, convert) data types
  - Numeric variables should be **integer** or **numeric**
  - Date-time variables should be **Date** or **POSIXct**
  - Categorial variables should be **factor**
  - All others should be **character**
- Visualize distributions
  - Numeric variables: **histogram**
  - Categorical variables: **numerical summary**
  - Character variables: look at a small sample
- Is the data as expected?
  - Range and shape of distribution
  - Number and breakdown of categories
  - Properties of character variables
- Explain and/or fix discrepancies

Convert Messy Data to Tidy

CollAge	PtNumber	PtSex	Result
7	5143567	M	22
42	3459254	F	5
19	2332467	F	5
80	3445732	M	89
41	7245673	F	12

A data set is **tidy** if:

1. Each **variable** is in its own **column**
2. Each **case** is in its own **row**
3. Each **value** is in its own **cell**

Missing values

Count the number of `NA`s (missing values) in each column of `esr`.

Hide

```
esr %>%
    map_df(function(x) sum(is.na(x))) %>%
    gather(column, NAs)
```

column <chr>	NAs <int>
CollAge	0
CollectDateTime	0
PtName	0
PtNumber	0
PtSex	0
Result	0
TstOrderName	0
UserQAFlags	0

8 rows

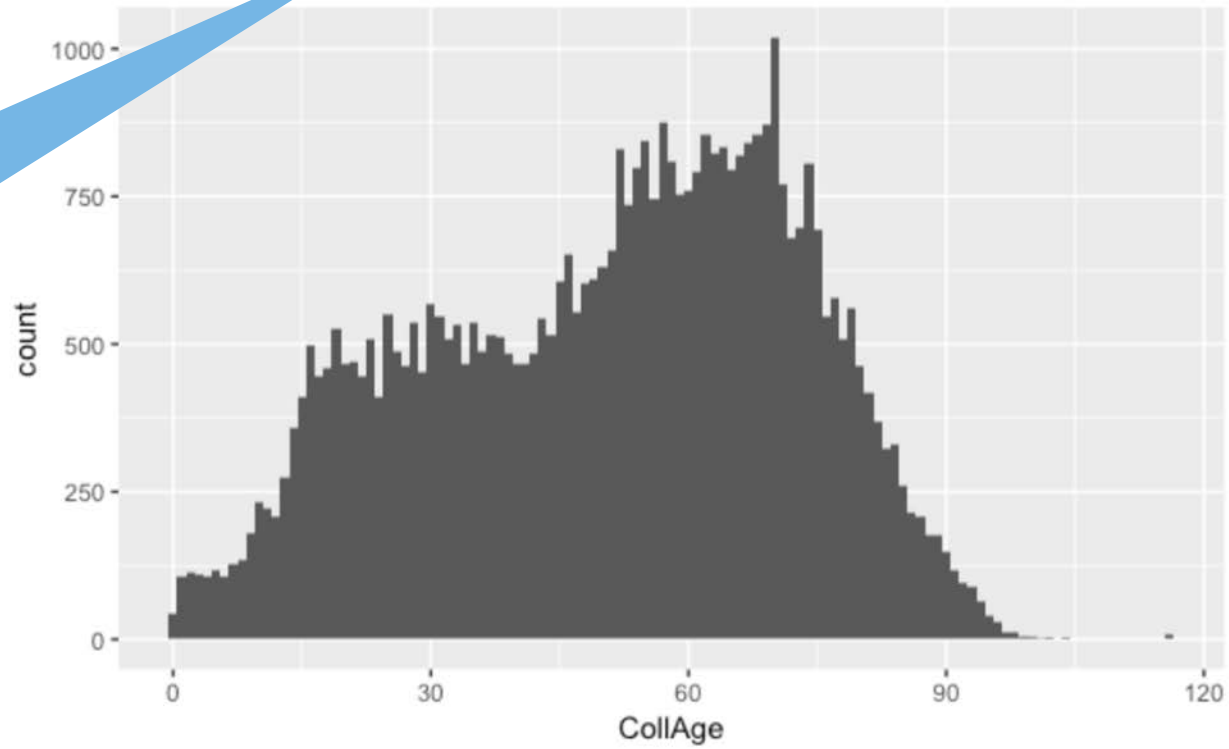
None of the columns of `esr` contain any missing values.

CollAge

`CollAge`, the age at collection, is an **integer** column, which is appropriate.

Hide

```
ggplot(data = esr) +
    geom_histogram(mapping = aes(x = CollAge),
                   binwidth = 1)
```



`binwidth`: width of bins. Experiment!

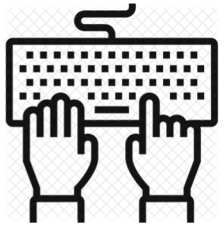
The distribution of `CollAge` is bimodal, with peaks at around 30 and at around 70 years.

CollectDateTime

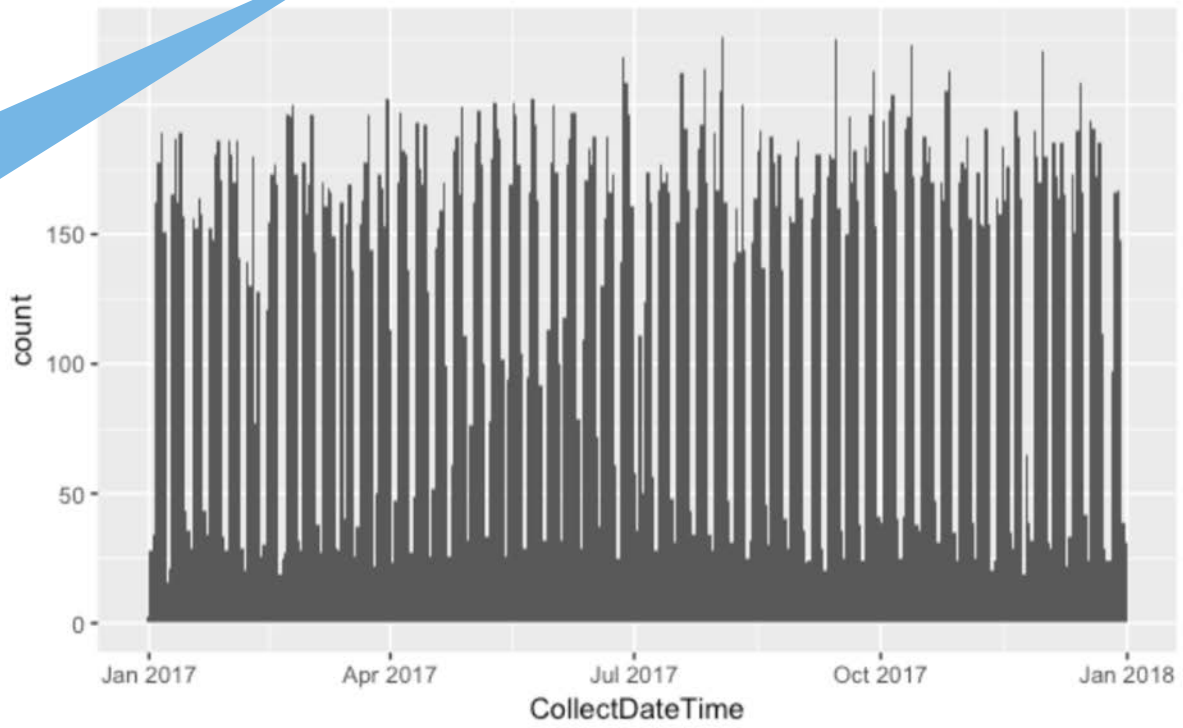
`CollectDateTime` is a **POSIXct** (datetime) column, which is appropriate.

Hide

```
ggplot(data = esr) +
    geom_histogram(mapping = aes(x = CollectDateTime),
                   bins = 365)
```



bins:
total number of bins.



Weekly cycling of ESR test volume is apparent. The weekly ESR test volume was approximately constant throughout 2017, without clear peaks or troughs.

Your Turn #3

Do you agree that weekly ESR testing volume was constant?

Can you think of an easy graph to verify this idea?

Discuss with your group and then try to create the graph in the R Markdown document.

set.seed():
set the random number generator
makes sure "random" results are reproducible

sample_n():
sample rows from a data frame

pull():
extract a single column and display it compactly

PtName

`PtName` is a **character** column, which is appropriat

Hide

```
set.seed(1)
esr %>%
    sample_n(10) %>%
    pull(PtName)
```

```
[1] "K[redacted],[redacted]"                "C[redacted],L[redacted]"
[3] "M[redacted],C[redacted]"               "M[redacted] S[redacted],I[redacted]"
[5] "B[redacted],J[redacted]"               "M[redacted],J[redacted]"
[7] "C[redacted],W[redacted] E[redacted]"   "C[redacted],R[redacted]"
[9] "D[redacted],D[redacted]"               "D[redacted],R[redacted]"
```

A random sample of 10 names is as expected.

PtNumber

`PtNumber`, the MGH MRNs of the patients, is a **character** column, which is appropriate.

Note: if `PtNumber` were converted to integer, MRNs with leading zeroes would be altered.

Hide

```
set.seed(1)
esr %>%
    sample_n(10) %>%
    pull(PtNumber)
```

```
 [1] "[redacted]" "[redacted]" "[redacted]" "[redacted]" "[redacted]" "[redacted]" "[redacted]"
 [8] "[redacted]" "[redacted]" "[redacted]"
```

A random sample of 10 MRNs shows that all are seven digits long, as expected for MGH patients.

PtSex

`PtSex` is a **character** column, but since this is a categorial variable, we will convert it to a **factor**.

Hide

```
esr <- esr %>%
    mutate(PtSex = as_factor(PtSex))

esr %>%
  pull(PtSex) %>%
  summary()
```

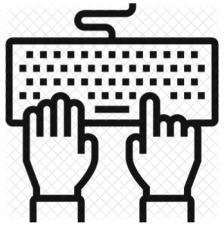
```
    F     M     U
26701 19857    17
```

The majority of ESR tests were performed on female (`F`) patients. Unexpectedly, in addition to `M` and `F` , there are some rows with `PtSex` denoted as `U` .

summary(): concise statistical summary

as_factor(): convert to factor

Overwrite existing PtSex variable



Overwrite existing esr data frame

```
esr %>%
    filter(PtSex == "U")
```

CollAge <int>	CollectDateTime <S3: POSIXct>	PtName <chr>	PtNumber <chr>	PtSex <fctr>
35	2017-[redacted]	[redacted]	[redacted]	U
39	2017-06-01 09:20:00	MGW CAP,ESR 01	QAPR-7475	U
33	2017-06-01 09:21:00	MGW CAP,ESR 02	QAPR-7476	U
35	2017-06-01 09:21:00	MGW CAP,ESR 03	QAPR-7477	U
35	2017-[redacted]	[redacted]	[redacted]	U

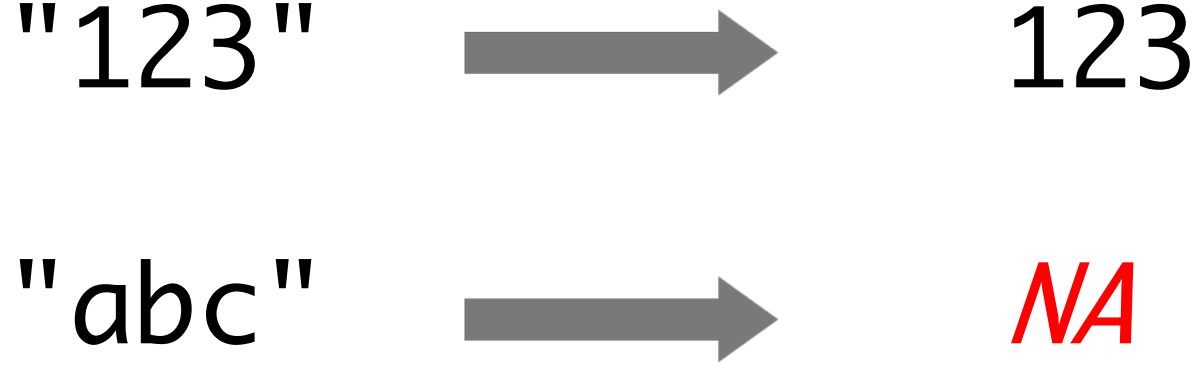
1-10 of 17 rows | 1-5 of 8 columns Previous 1 2 Next

It appears that the rows with `PtSex` equal to `U` belong to a small number of patients as well as AP survey samples. Since these rows represent less than 0.1% of the whole data set, it will be s to remove them.

```
esr <- esr %>%
    filter(PtSex != "U")
```

Conversion of Character to Numeric Types can create Missing Values!

`as.integer()`



get non-numeric values of `Result`, arranged by order of frequency

convert `Result` to integer and remove all rows where `Result` is now *NA*

Result

`Result` is a **character** column but should be an **integer**, so we will convert it.

Before doing so, examine any rows that do *not* contain an integer value:

```
esr %>%
    filter(is.na(as.integer(Result))) %>%
    group_by(Result) %>%
    summarize(n = n()) %>%
    arrange(desc(n))
```

Result <chr>	n <int>
REFUS	1292
CREDIT	118
>140	14
QNS	12
CANCEL	6

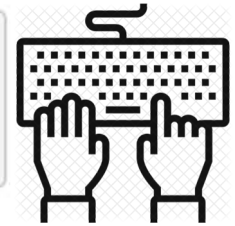
1-10 of 19 rows Previous **1** 2 Next

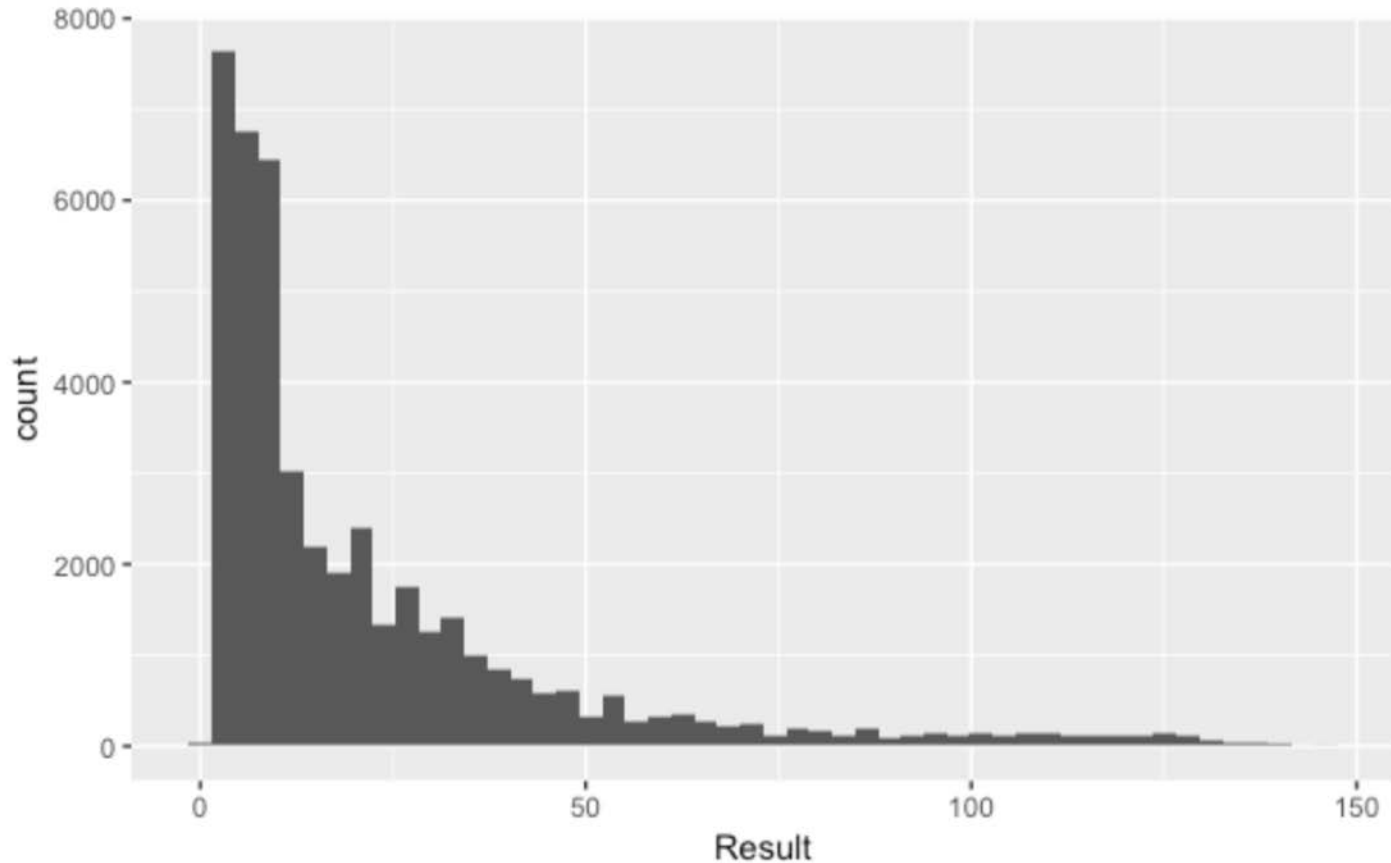
About 1400 rows (out of about 47,000) have a non-integer value in `Result` - about **3%**. The large majority of them have a value of `REFUS` or `CREDIT`, indicating that the sample was not run or reported. These rows should be discarded.

Only a very small number of samples had excessive values (`>140`, `>144`). It is acceptable to discard these rows as well.

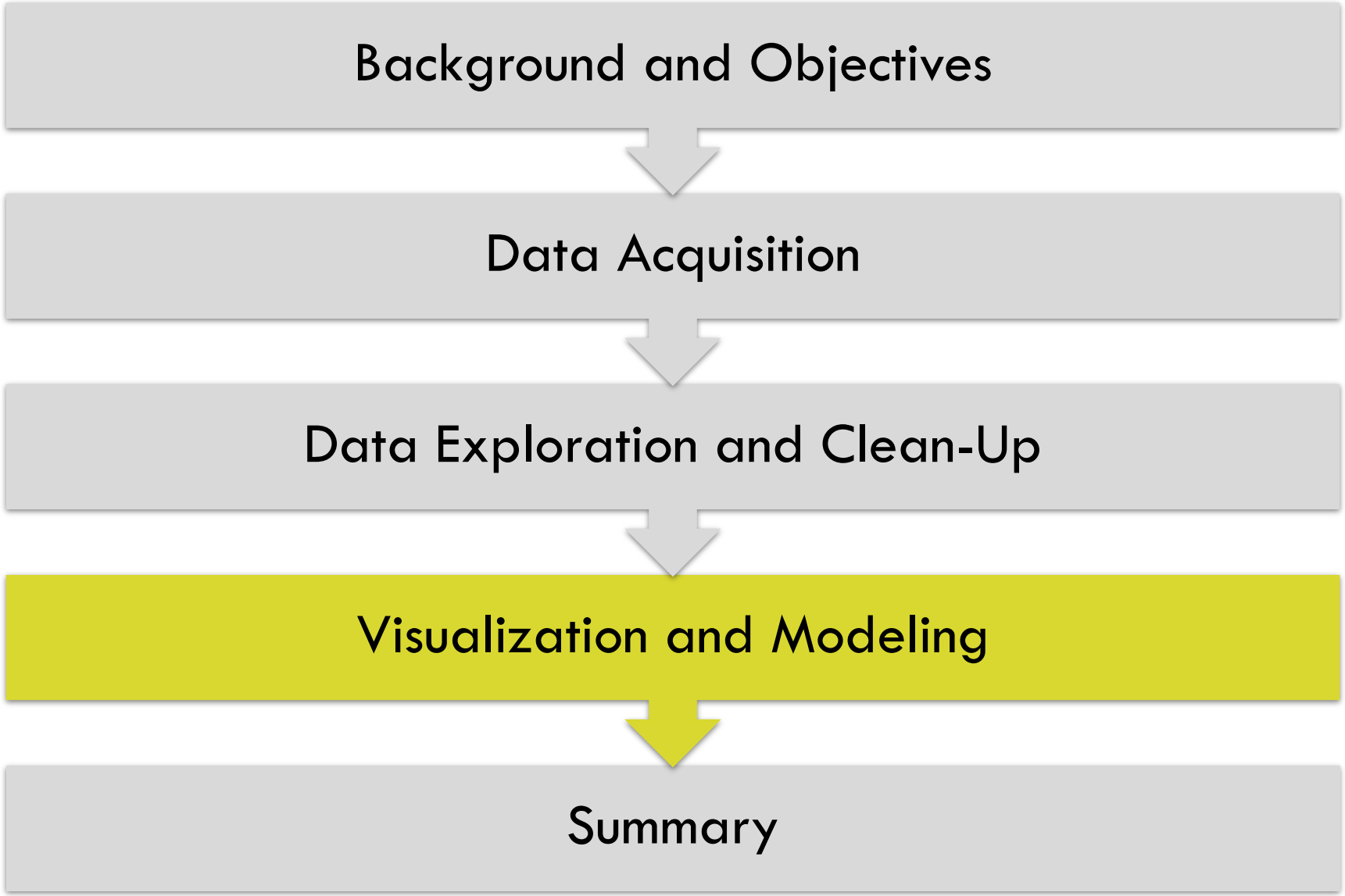
```
esr <- esr %>%
    mutate(Result = as.integer(Result)) %>%
    drop_na()
```

```
ggplot(data = esr) +
    geom_histogram(mapping = aes(Result),
                   bins = 50)
```





The distribution of ESR values appears to have a mode at around 10, with a long right tail, which represents the abnormal values.


Background and Objectives
Data Acquisition
Data Exploration and Clean-Up
Visualization and Modeling
Summary

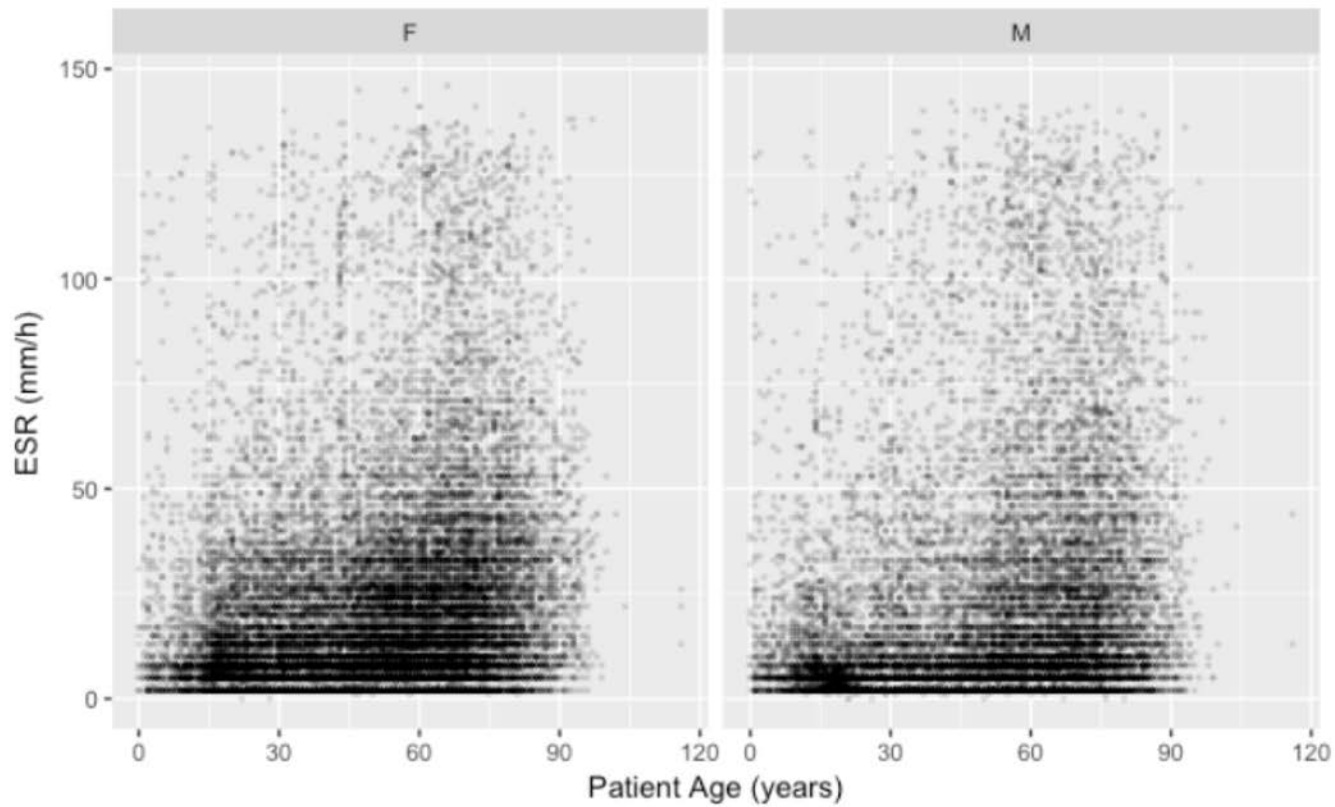
Visualization and Modeling

Iterative cycles of:

- Pose a question or hypothesis.
- Create a graph or statistical summary that addresses the question or tests the hypothesis.
- Comment on new insights gained and new problems raised.

The primary goal of this project is to test is whether normal and abnormal ESR values can be separated into distinct populations, so that reference intervals can be set at the boundary. From the literature, I expect that ESR values increase with age and sex. To reduce these confounding effects, we will look at the relationship of ESR and age, broken down by sex.

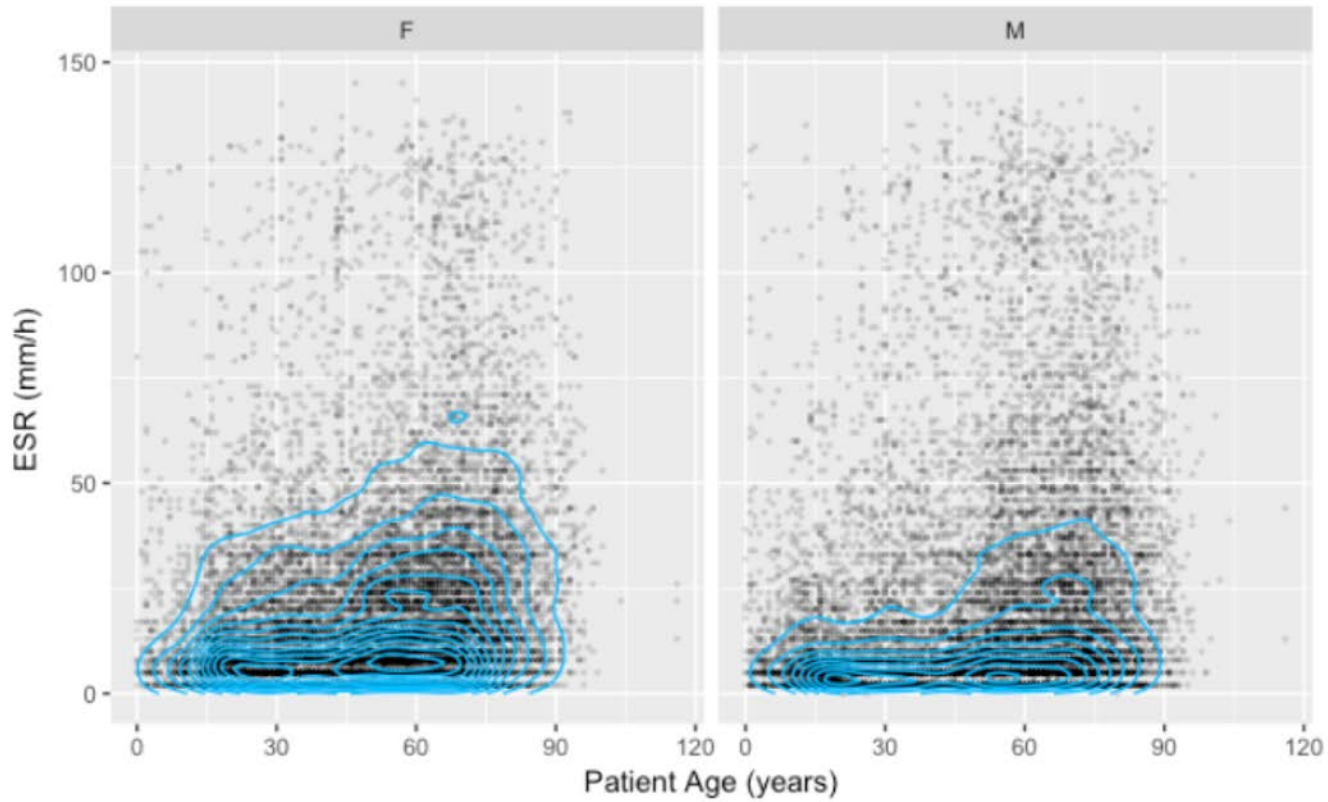
Code



There is no clear separation of normal and abnormal values. In addition, there are more data points on the `female` panel, which makes this graph difficult to interpret.

Will a separation become apparent when equal numbers of points are shown on each graph, and a density contour plot is overlaid?

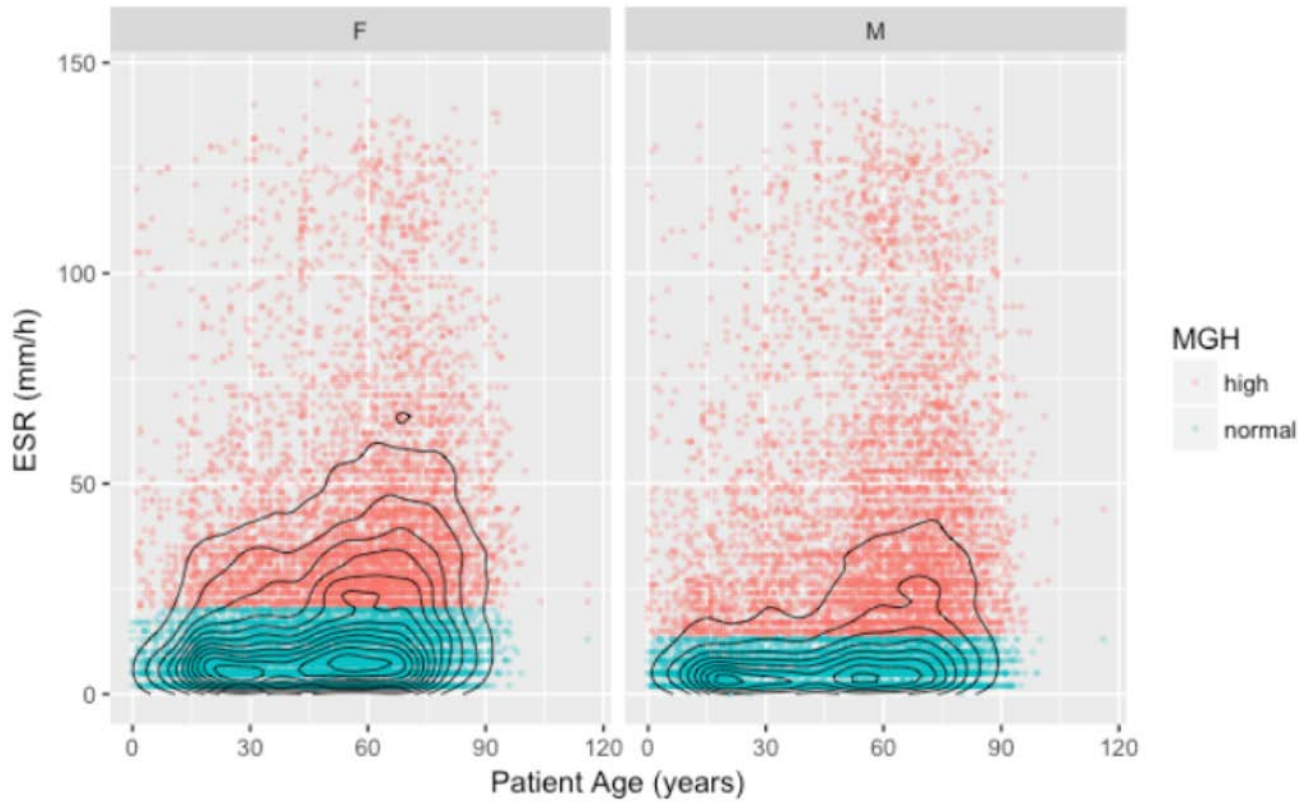
Code



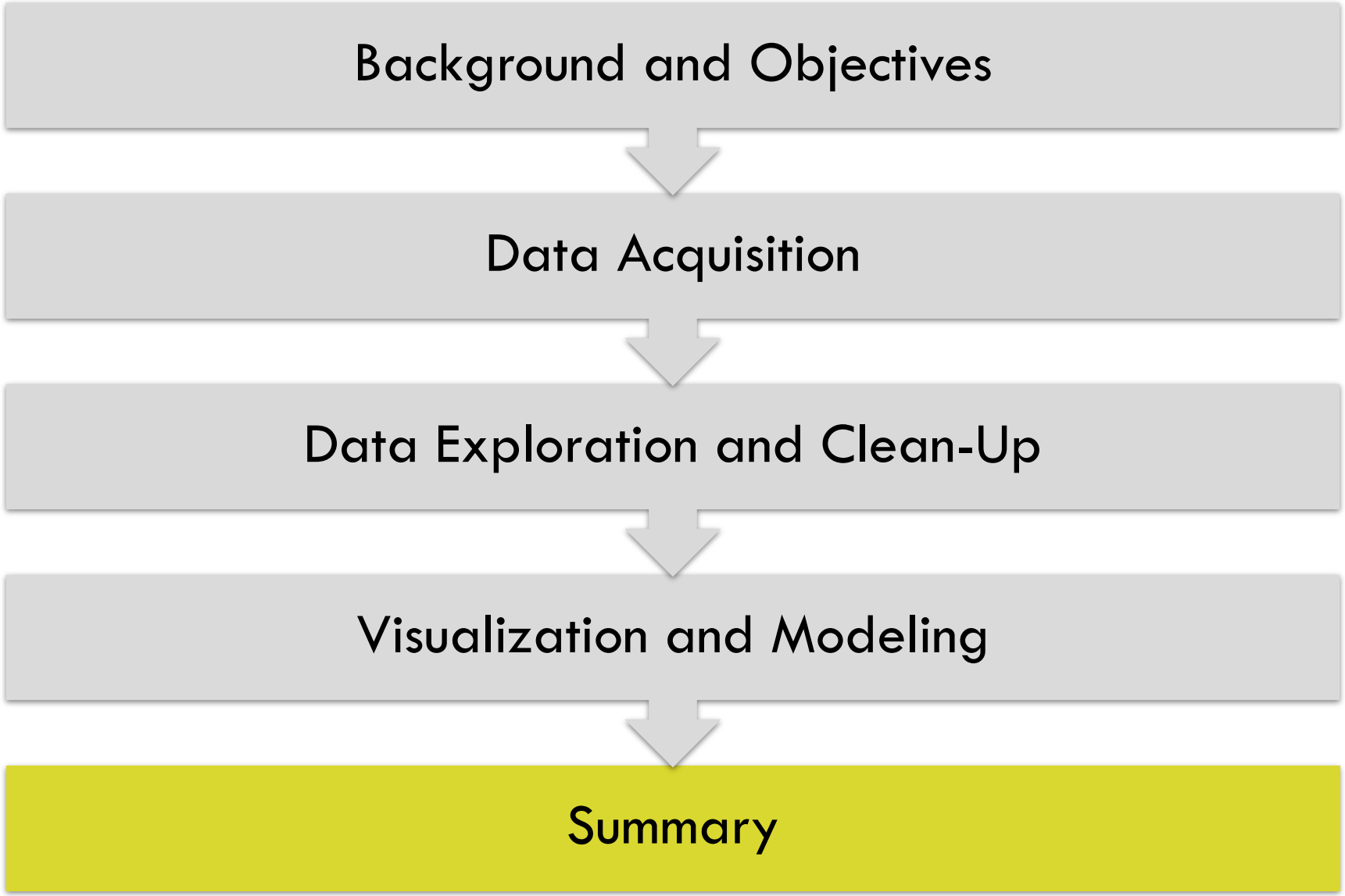
Subsampling and density contour plotting did not reveal separable populations. This graph clearly shows that female patients tended to have higher ESR levels than male patients, and that older patients tended to have higher ESR levels than younger patients. However, this effect was not universal, and many old patients of both genders had low ESR values.

Do the MGH reference ranges appear to appropriately capture patients with normal ESRs? Abnormal ESRs?

Code



MGH's normal reference range appears to capture the majority of normal ESR results. It does not capture the tendency of older individuals to have higher ESR values. However, this tendency is not universal, and it is uncertain whether high ESR values in older individuals are physiologically normal.


Background and Objectives
Data Acquisition
Data Exploration and Clean-Up
Visualization and Modeling
Summary

Summary

- Briefly restate background, objectives, and data acquisition strategy
- Restate each important result
- Explain how the new insights will be used to improve patient care

Your Turn #4

How would you summarize the ESR report?
Discuss with your group mates.