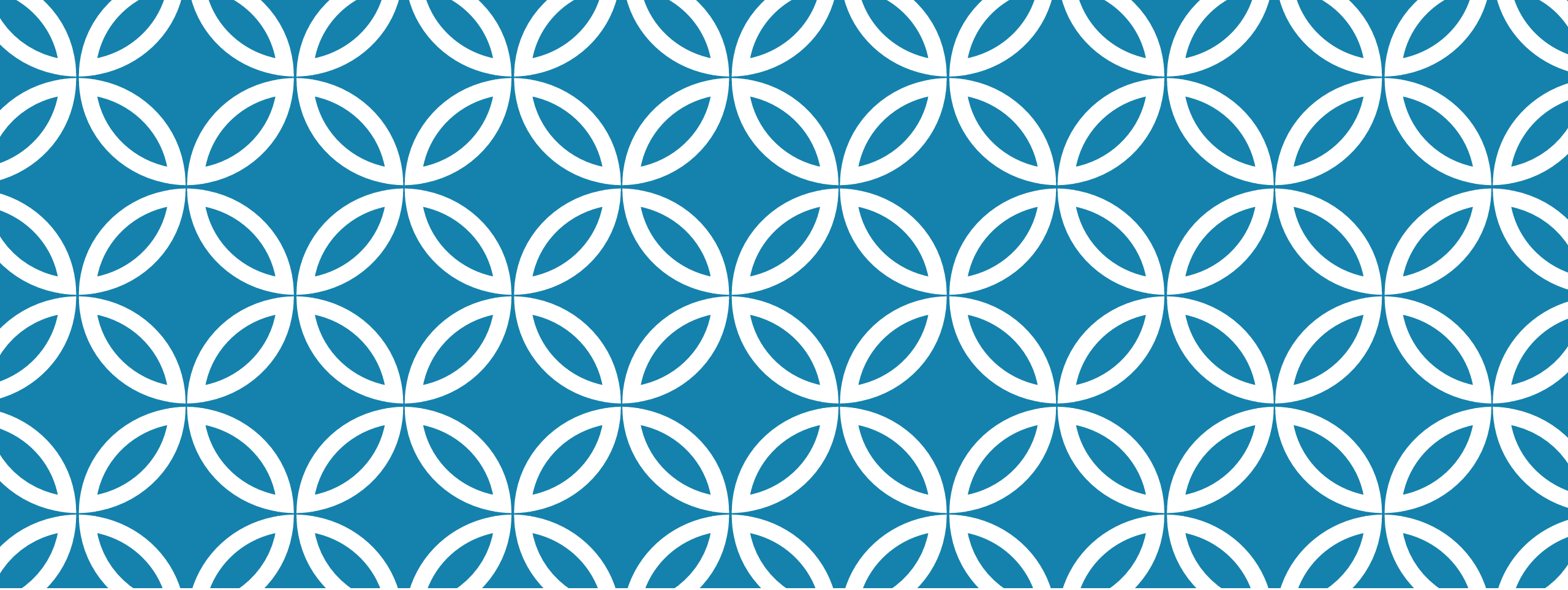


Exploratory Data Analysis in R

Session 6

Amrom Obstfeld

Introduction to R Workshop

May 6, 2019

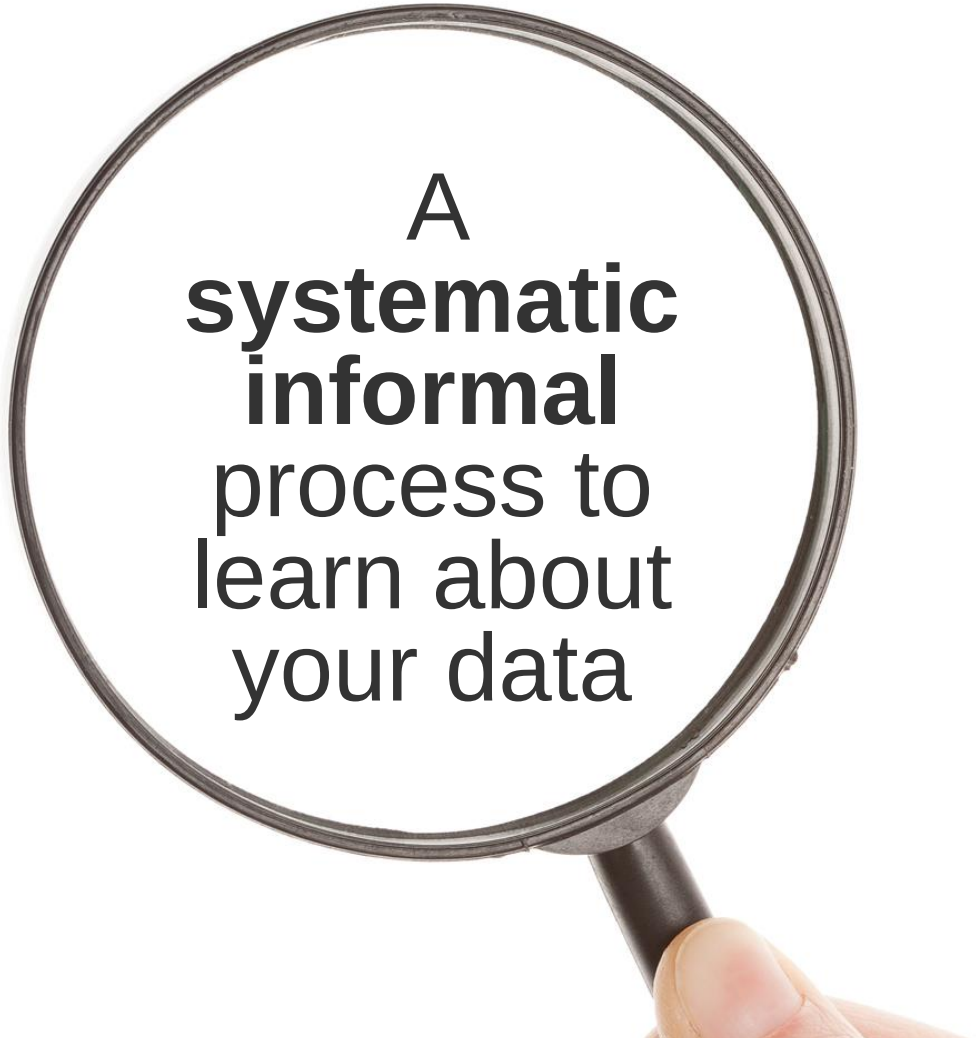
7:00 am–8:00 am	BREAKFAST BALLROOM LOBBY 2ND FLOOR
8:00 am–8:10 am	Instructor and Course Introduction
8:10 am–9:50 am	Introduction to R and RStudio for Reproducible Reporting
9:50 am–10:10 am	REFRESHMENT BREAK BALLROOM LOBBY 2ND FLOOR
10:10 am–11:50 am	Data Wrangling
12:00 pm–1:00 pm	LUNCH BALLROOM LOBBY 2ND FLOOR
1:00 pm–2:50 pm	Data Understanding
2:50 pm–3:10 pm	REFRESHMENT BREAK BALLROOM LOBBY 2ND FLOOR
3:10 pm–5:00 pm	Exploratory Data Analysis

Goals and Objectives

- Appreciate the role and process of exploratory data analysis (EDA) in understanding data
- Learn about MIMIC-III data set and its utility in learning to work with biomedical data
- Further reinforce the skills learned during the previous five sessions

What is EDA

Exploratory Data Analysis



Exploratory Data Analysis

EDA is an iterative process in which we:

1. Generate questions about our data.
2. Search for answers by visualizing, transforming, and modelling our data.
3. Use what we learn to refine your questions and/or generate new questions.

Exploratory Data Analysis

Two Fundamental Questions

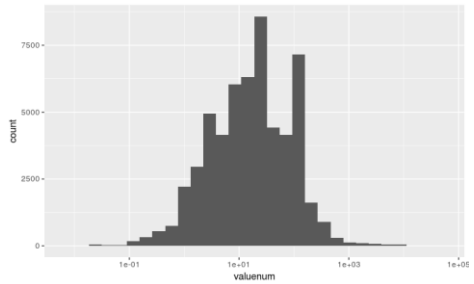
1. What is the distribution of data in each of my variables
2. How do my variables relate with one another

How do we explore variation in data?

Quantitative

Range, mean, median, mode

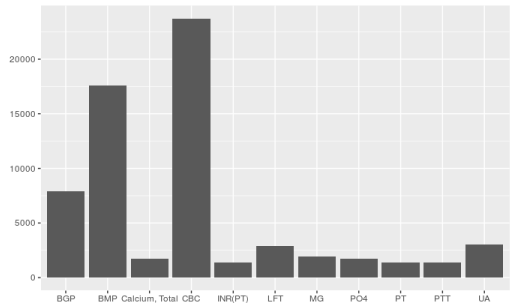
Histograms



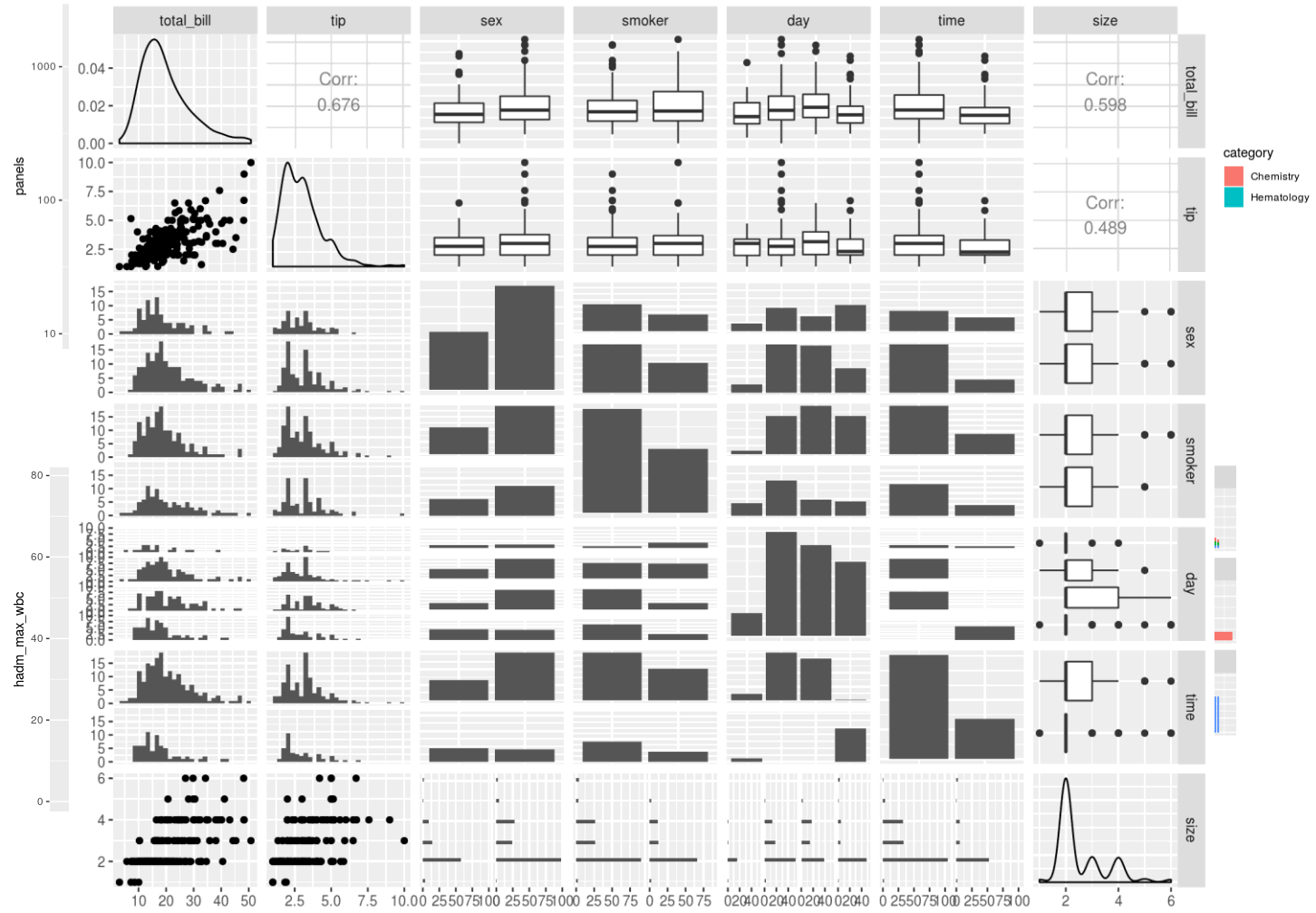
Categorical

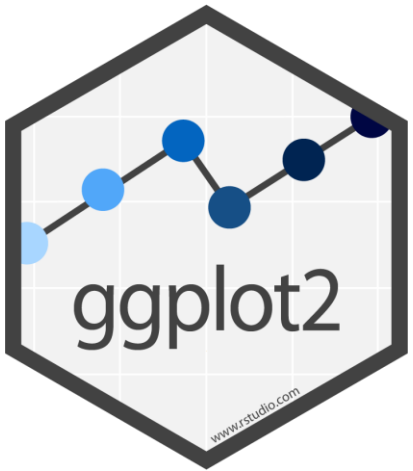
Frequency (counts, percent)

Bar chart



How do we explore COvariation in data?





EDA as Data QC

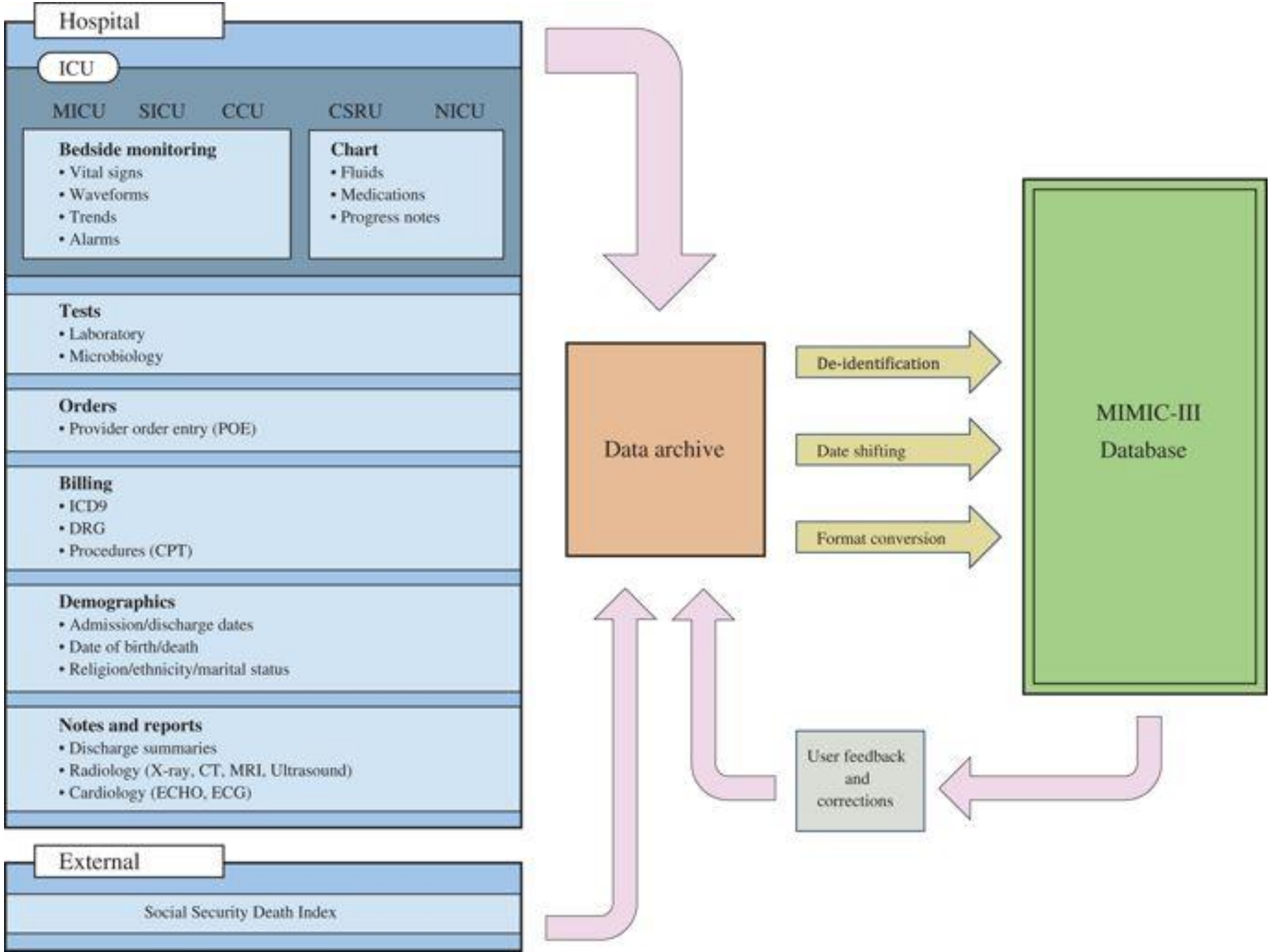
Investigate:

- Missing data
- Outliers
- Logical discrepancies

Medical Information Mart for Intensive Care (MIMIC)-III data set

MIMIC Data set

- MIMIC-III is a widely-used, freely available biomedical dataset
- Developed by the MIT Lab for Computational Physiology
- Deidentified health data associated with >40,000 critical care patients
- Includes demographics, vital signs, laboratory tests, medications, free text notes and more
- Details are available at https://mimic.physionet.org/


Hospital
ICU
MICU
SICU
CCU
CSRU
NICU
Bedside monitoring
• Vital signs
• Waveforms
• Trends
• Alarms
Chart
• Fluids
• Medications
• Progress notes
Tests
• Laboratory
• Microbiology
Orders
• Provider order entry (POE)
Billing
• ICD9
• DRG
• Procedures (CPT)
Demographics
• Admission/discharge dates
• Date of birth/death
• Religion/ethnicity/marital status
Notes and reports
• Discharge summaries
• Radiology (X-ray, CT, MRI, Ultrasound)
• Cardiology (ECHO, ECG)
External
Social Security Death Index
Data archive
De-identification
Date shifting
Format conversion
MIMIC-III
Database
User feedback
and
corrections

MIMIC Data Access

- Course in protecting human research participants including HIPAA requirements.
- Data use agreement
  - outlines appropriate data usage
  - security standards
  - forbids efforts to identify individual patients

Common uses for MIMIC

- Educational coursework
  - medical analytics courses
- Research
  - machine learning approaches for prediction of patient outcomes
  - semantic analysis of unstructured patient note
- Datathons

EDA Exercise

EDA exercise outline

For each exercise:

- Explicit task description
- Use functions to complete the task
- Question-set based on your findings

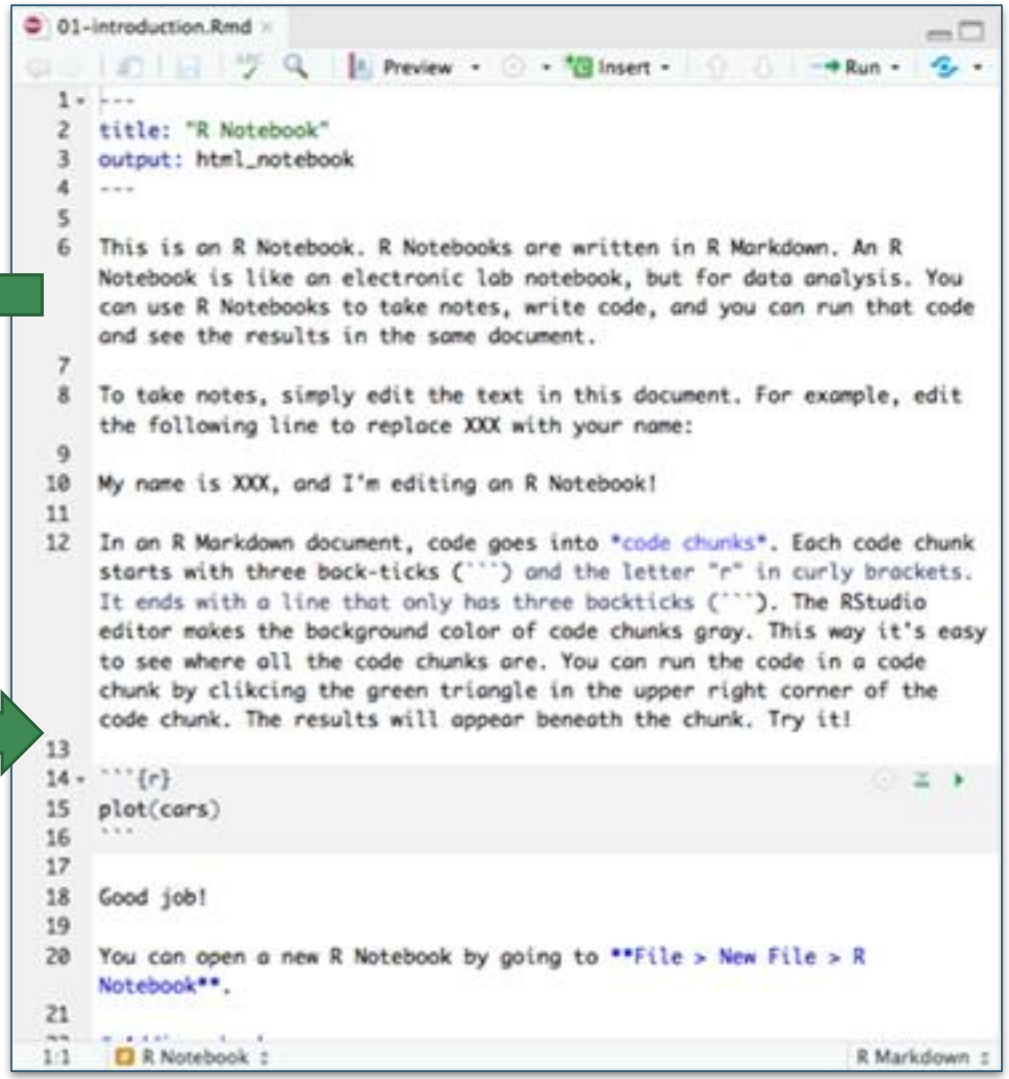
EDA exercise outline

Your Turn

Introduce yourself to your neighbors

- Who are you?
- Where are you from?
- What do you do with data?
- Have you ever used R?

3:00



Exercise 0

Exercise 0

Use the code block below to run `library()` on "tidyverse"

Exercise 1

Use the read_csv() function to read mimic.csv into a new data frame called "mimic"

Explore the data frame using some of the tools we've learned today (summary(), head())

1. How many rows are in the data frame?
2. How many columns are in the data frame?
3. What does each row in the data frame represent?
4. How many columns contain information about the patient's admission and how many relate to the test order?

Exercise 2

- Use `filter()` to find the rows that have NA in the valuenum column
- Use `select()` to narrow down to just the "panel_test", "test_name", "component", "value" and "valuenum" columns
- Use `arrange()` to order the data frame by "value" and "component" columns

```
mimic %>%
    filter(is.na(valuenum) %>%
    select(panel_test, component,
                  value, valuenum) %>%
    arrange(value,component)
```

1. What is the difference between the "value" and "valuenum" columns?
2. What kind of result values in the data set appear in the "value" column but are NA in the "valuenum" column?

Exercise 3

- Use group_by() and summarize() to get a sense for the counts of data in some of the columns with categorical data
- Use n() and n_distinct inside of the summarize() function to count the rows and distinct values in each categorical variable

```
mimic %>%
    group_by(religion) %>%
    summarise(d_pt =  n_distinct(subject_id))
```

1. How many distinct patients and admissions are in the data frame?
2. How many different panel tests and components are in the data frame?
3. What is the most common religion for patients in the data frame?
4. *Challenge Question* What is the most commonly ordered test in the data set?

Exercise 4

- Use ggplot() to assess the distribution of charttime and valuenum
- Try different scales to visualize the laboratory results
- Use the fill aesthetic to parse out the contribution of different categorical variables, such as category or fluid types, to the distribution of laboratory values

```
ggplot(mimic)+
  geom_histogram(aes(valuenum,fill=category))+
  scale_x_log10()
```

1. What do you notice about the pattern of charttimes and valuenums in the data frame?
2. Are there differences between the distribution of results for "Chemistry" and "Hematology" test categorys?
3. Are there outlier categories with only a few results?
4. *Challenge Question* Can you estimate what the reference range is for the "Hemoglobin"?

Show and Tell 1 – What about NAs?

- Missing data (NA) is important to be aware of because its presence may indicate a problem with the data or may influence the statistics and conclusions we make

```
mimic %>%
    mutate_all(is.na) %>%
    summarise_all(sum) %>%
    arrange(desc(value))
```

Show and Tell 2 – Characters in numeric data

- Analyzing laboratory data can be challenging for several reasons
- One reason is that numeric data can often be stored as characters

```
mimic %>%
    filter(str_detect(value,">|<")) %>%
    mutate(valuenum =
           ifelse(str_detect(value,">|<"),
                  parse_number(value),
                  valuenum)) %>%
    select(test_name, component, value,
valuenum)
```

Show and Tell 3 – Working With Dates

- Medical data is inherently temporal in nature
- R has several packages that help in dealing with datetime based data

```
library(lubridate)
mimic %>%
   mutate(charthour = hour(charttime)) %>%
   filter(!category %in%
          c("HEMATOLOGY","CHEMISTRY")) %>%
   distinct(curr_service, category, charttime,
            charthour, hadm_id, panel_test) %>%
ggplot(aes(x=charthour,fill=category))+
  geom_bar()+
  facet_wrap(~curr_service, scales = "free_y")
```