Introduction to Medical Statistics 2026

Oxford University Clinical Research Unit
March 23-27, 2026

Ronald Geskus and the biostatistics crew
Oxford University Clinical Research Unit
Hospital for Tropical Diseases, Ho Chi Minh City, Viet Nam

Part II

Exploratory Data Analysis

Contents:

1. (Baseline) tables
2. Graphs: $\geq$ 2 categorical variables
3. Graphs: numeric variables by group
4. Graphs: 2 numeric variables

Outline

Exploratory Data Analysis (EDA)

Graphs: two or more categorical variables

Numeric variable by groups

Two numeric variables

Data visualization in R

Wikipedia (March 2026)

Exploratory data analysis (EDA) is an approach of analyzing data sets to summarize their main characteristics, often using statistical graphics and other data visualization methods. A statistical model can be used or not, but primarily EDA is for seeing what the data can tell beyond the formal modeling and thereby contrasts with traditional hypothesis testing, in which a model is supposed to be selected before the data is seen. Exploratory data analysis has been promoted by John Tukey since 1970 to encourage statisticians to explore the data, and possibly formulate hypotheses that could lead to new data collection and experiments. EDA is different from initial data analysis (IDA), which focuses more narrowly on checking assumptions required for model fitting and hypothesis testing, and handling missing values and making transformations of variables as needed. EDA encompasses IDA.

Exploratory Data Analysis (EDA)

1. Descriptive analysis (IDA): inspect and summarize data characteristics
   - variables: what type, which distribution
     Is there a need to transform numeric variable?
   - find errors, check for peculiarities, such as missing values and outliers

Exploratory Data Analysis (EDA)

1. Descriptive analysis (IDA): inspect and summarize data characteristics
   - variables: what type, which distribution
     Is there a need to transform numeric variable?
   - find errors, check for peculiarities, such as missing values and outliers
2. EDA: find patterns and relationships between variables

Exploratory Data Analysis (EDA)

1. Descriptive analysis (IDA): inspect and summarize data characteristics
   - variables: what type, which distribution
     Is there a need to transform numeric variable?
   - find errors, check for peculiarities, such as missing values and outliers
2. EDA: find patterns and relationships between variables
   - numerical summary of variables by subgroup

Baseline table

- Most papers include a numerical description of all variables that are relevant for a study (often in “Table 1”)
- Such summaries often reported by subgroup
  Example: Recovery Trial Dexamethasone in Covid-19 patients

DEXAMETHASONE IN HOSPITALIZED PATIENTS WITH COVID-19

Table 1. Characteristics of the Patients at Baseline, According to Treatment Assignment and Level of Respiratory Support.*

Characteristic	Treatment Assignment		Respiratory Support Received at Randomization		
	Dexamethasone (N=2104)	Usual Care (N=4321)	No Receipt of Oxygen (N=1535)	Oxygen Only (N=3883)	Invasive Mechanical Ventilation (N=1007)
Age†					
Mean — yr	66.9±15.4	65.8±15.8	69.4±17.5	66.7±15.3	59.1±11.4
Distribution — no. (%)					
<70 yr	1141 (54)	2504 (58)	659 (43)	2148 (55)	838 (83)
70 to 79 yr	469 (22)	859 (20)	338 (22)	837 (22)	153 (15)
≥80 yr	494 (23)	958 (22)	538 (35)	898 (23)	16 (2)
Sex — no. (%)					
Male	1338 (64)	2749 (64)	891 (58)	2462 (63)	734 (73)
Female‡	766 (36)	1572 (36)	644 (42)	1421 (37)	273 (27)
Median no. of days since symptom onset (IQR)§	8 (5–13)	9 (5–13)	6 (3–10)	9 (5–12)	13 (8–18)
Median no. of days since hospitalization (IQR)	2 (1–5)	2 (1–5)	2 (1–6)	2 (1–4)	5 (3–9)

Exploratory Data Analysis (EDA)

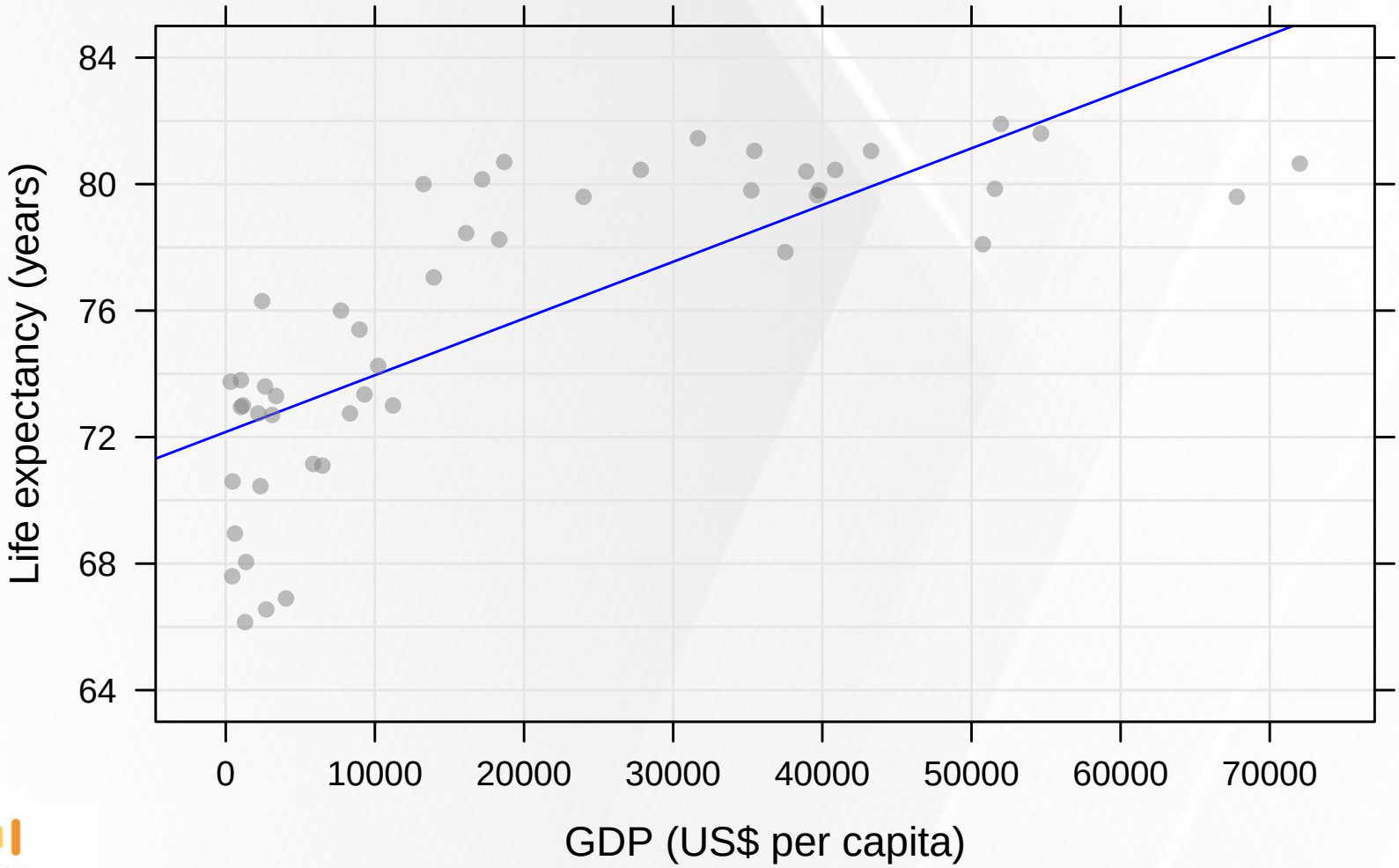
1. Descriptive analysis (IDA): inspect and summarize data characteristics
   - variables: what type, which distribution
     Is there a need to transform numeric variable?
   - find errors, check for peculiarities, such as missing values and outliers
2. EDA: find patterns and relationships between variables
   - numerical summary of variables by subgroup
3. Purpose EDA
   - suggest hypotheses for further research
   - suggest appropriate modeling approach
     example: WHO European Health Report data 2007

Exploratory Data Analysis (EDA)

1. Descriptive analysis (IDA): inspect and summarize data characteristics
   - variables: what type, which distribution
     Is there a need to transform numeric variable?
   - find errors, check for peculiarities, such as missing values and outliers
2. EDA: find patterns and relationships between variables
   - numerical summary of variables by subgroup
3. Purpose EDA
   - suggest hypotheses for further research
   - suggest appropriate modeling approach
     example: WHO European Health Report data 2007

WHO European Health Report data 2007

Linear trend not correct, more subtle pattern present



oucru

UNIVERSITY OF OXFORD

wellcome

Exploratory Data Analysis (EDA)

1. Descriptive analysis (IDA): inspect and summarize data characteristics
   - variables: what type, which distribution
     Is there a need to transform numeric variable?
   - find errors, check for peculiarities, such as missing values and outliers
2. EDA: find patterns and relationships between variables
   - numerical summary of variables by subgroup
3. Purpose EDA
   - suggest hypotheses for further research
   - suggest appropriate modeling approach
     example: WHO European Health Report data 2007
4. Often using graphics

   The greatest value of a picture is when it forces us to notice what we never expected to see (John W. Tukey)

Outline

Exploratory Data Analysis (EDA)

Graphs: two or more categorical variables

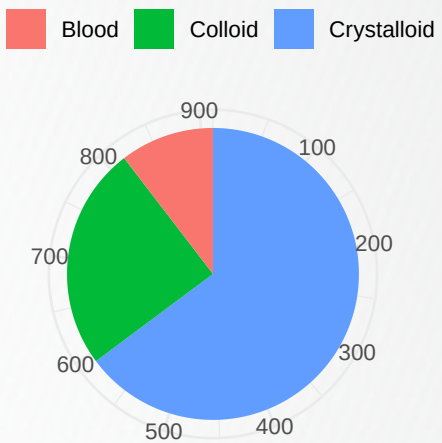
Numeric variable by groups

Two numeric variables

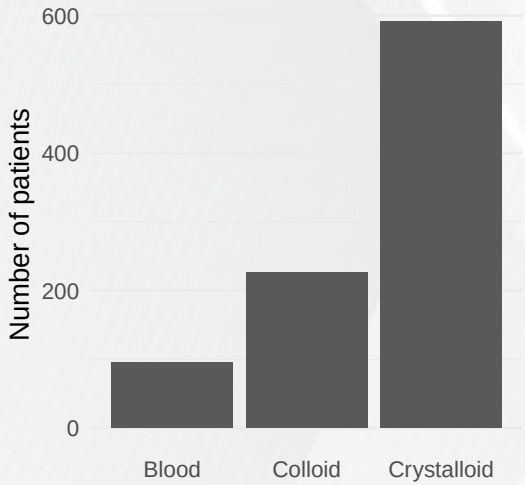
Data visualization in R

Main graph types for single categorical variable

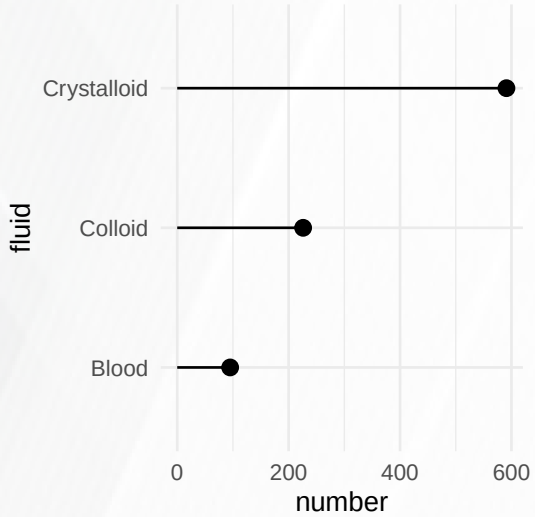
Pie chart



Barchart



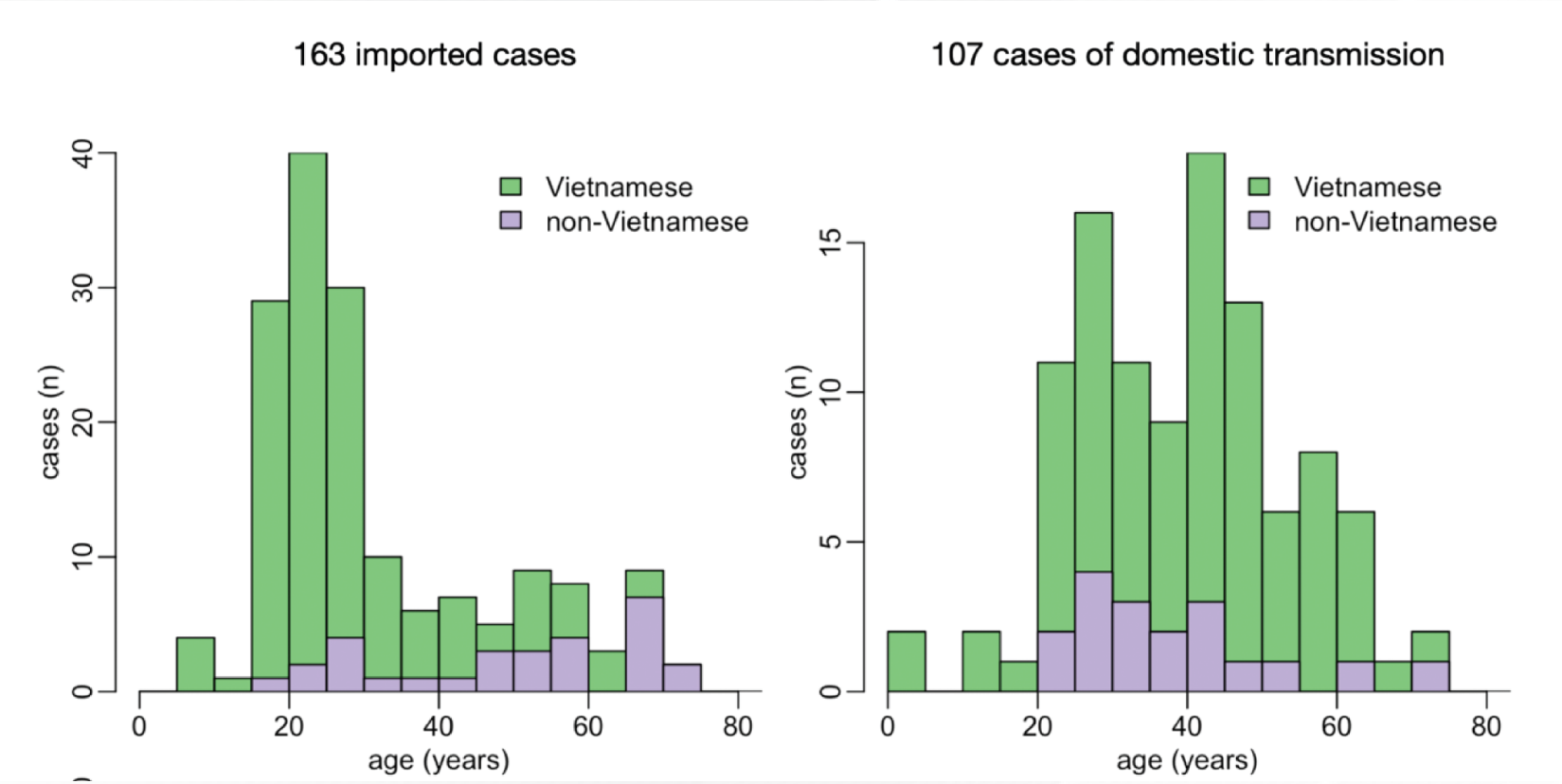
Dotplot



Summary by number or by percentage

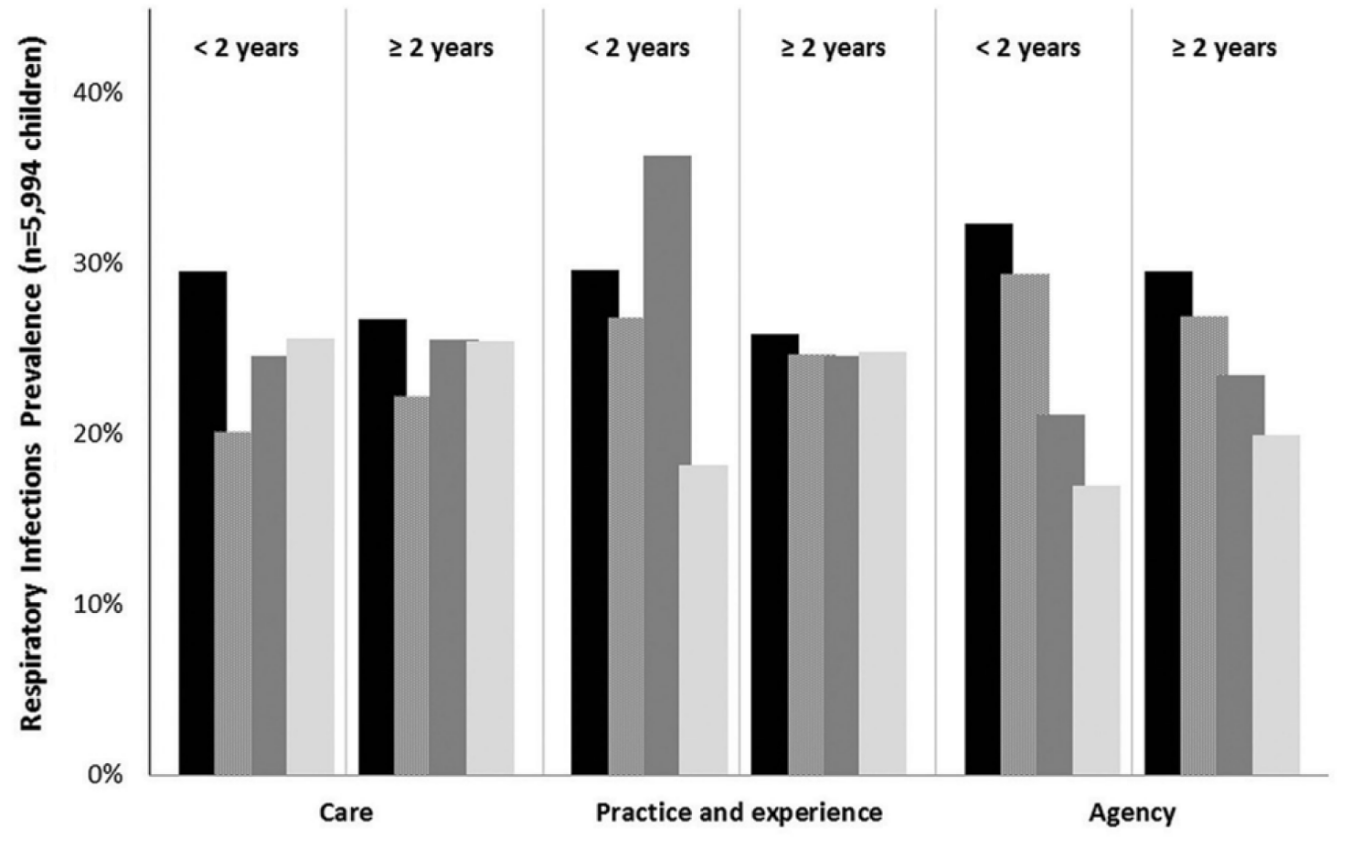
Stacked barchart (numbers; 2 facets/panels)

SARS-CoV-2 infections in Vietnam (January-April 2020)



Dodged barchart (percentages)

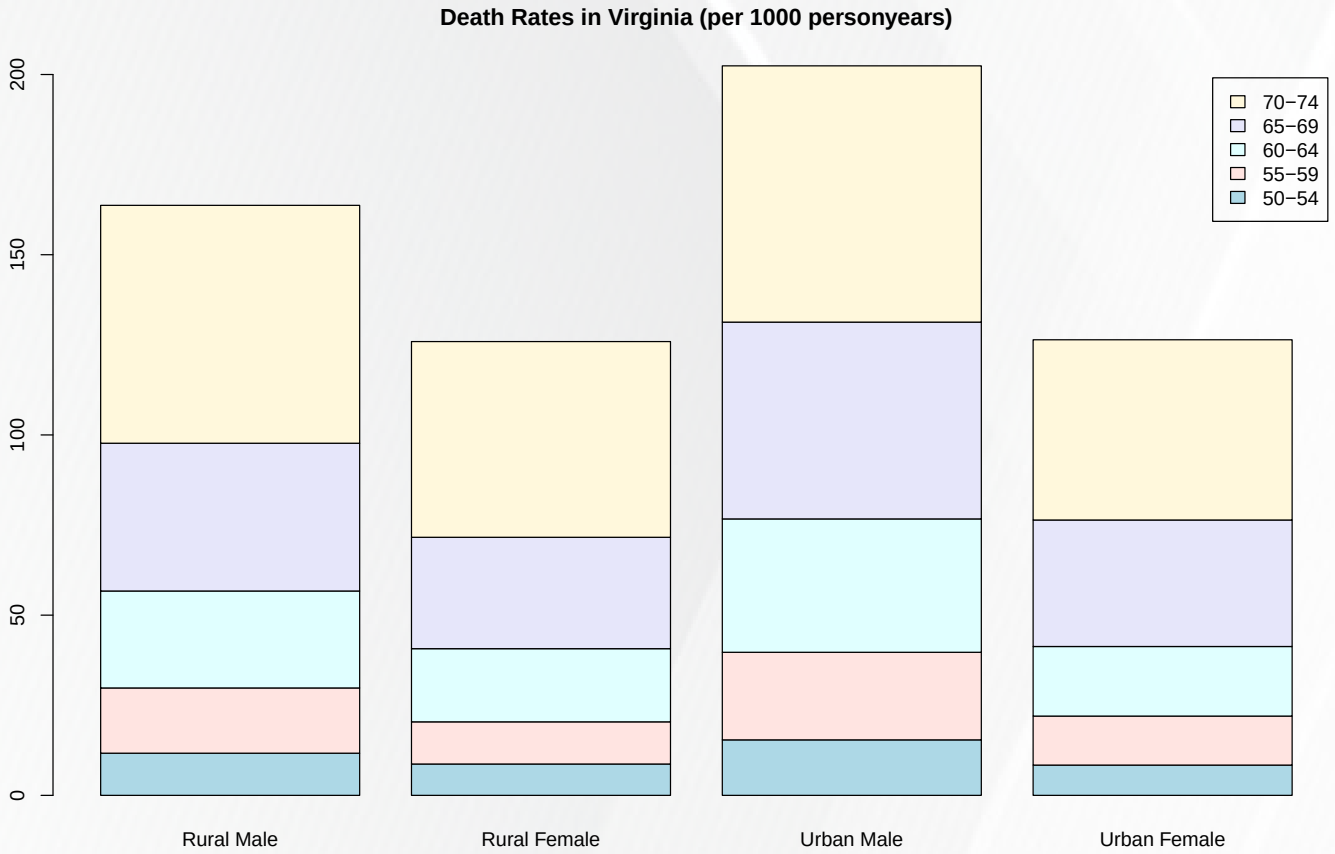
Prevalence respiratory infections by age group and 3 maternal factors (4 levels, low to high)



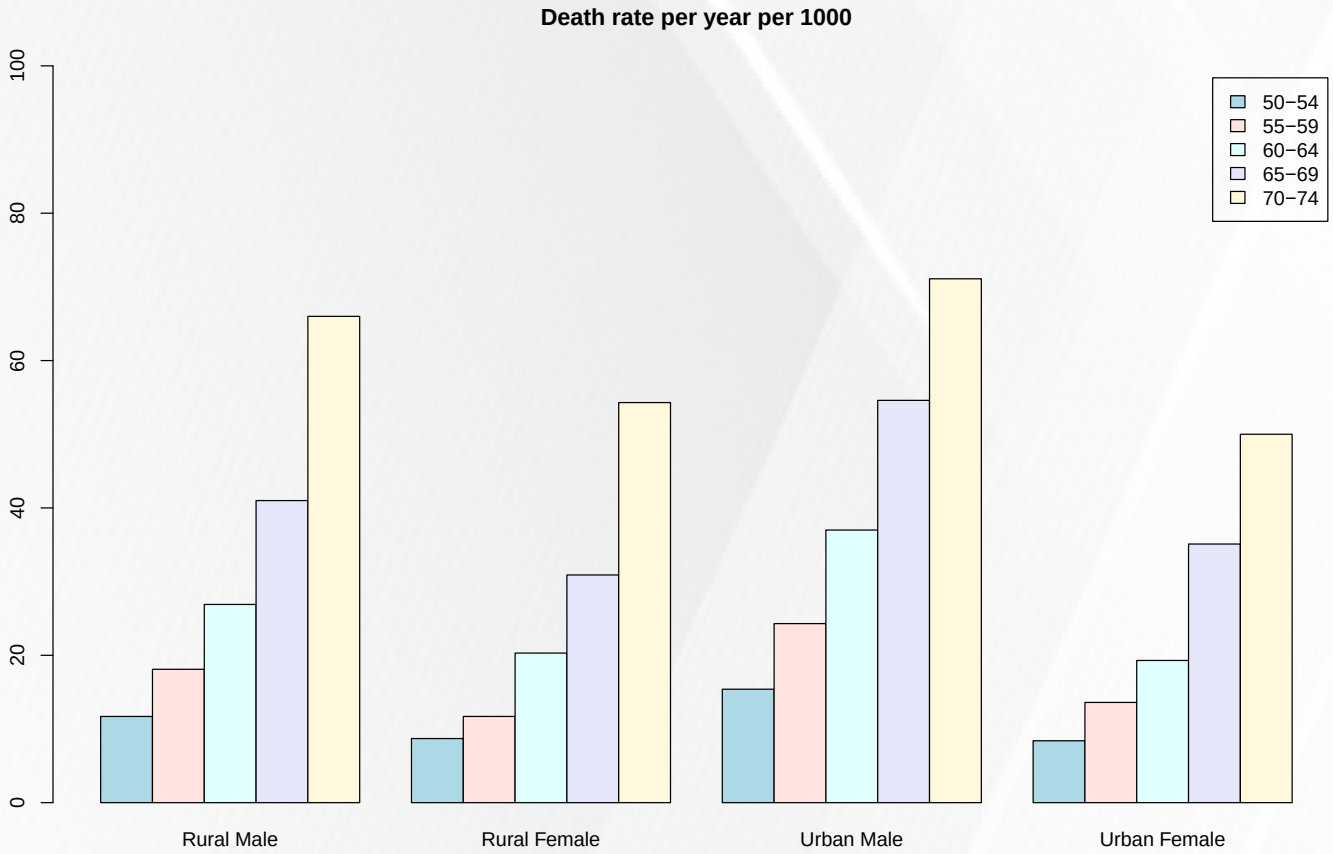
Mortality per 1000 inhabitants in Virginia in 1940

	Rural Male	Rural Female	Urban Male	Urban Female
50-54	11.70	8.70	15.40	8.40
55-59	18.10	11.70	24.30	13.60
60-64	26.90	20.30	37.00	19.30
65-69	41.00	30.90	54.60	35.10
70-74	66.00	54.30	71.10	50.00

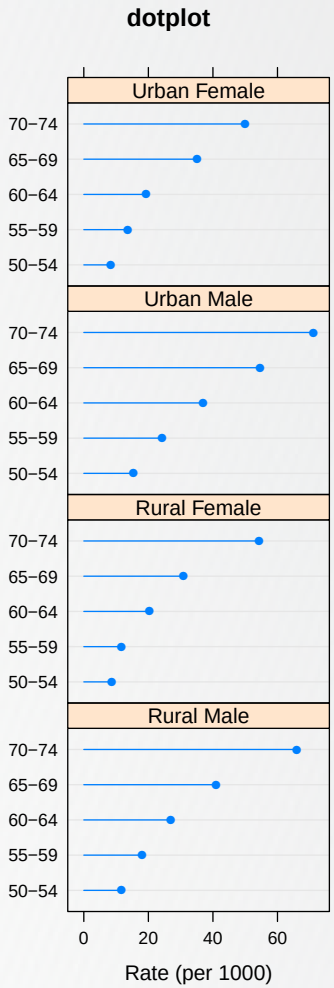
Stacked barchart

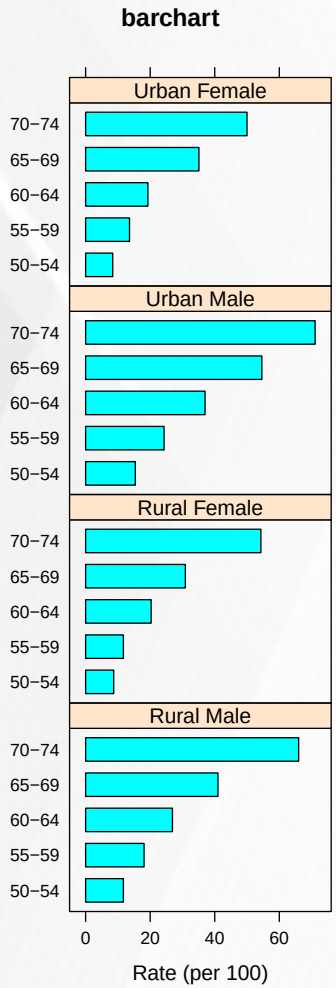


Dodged barchart

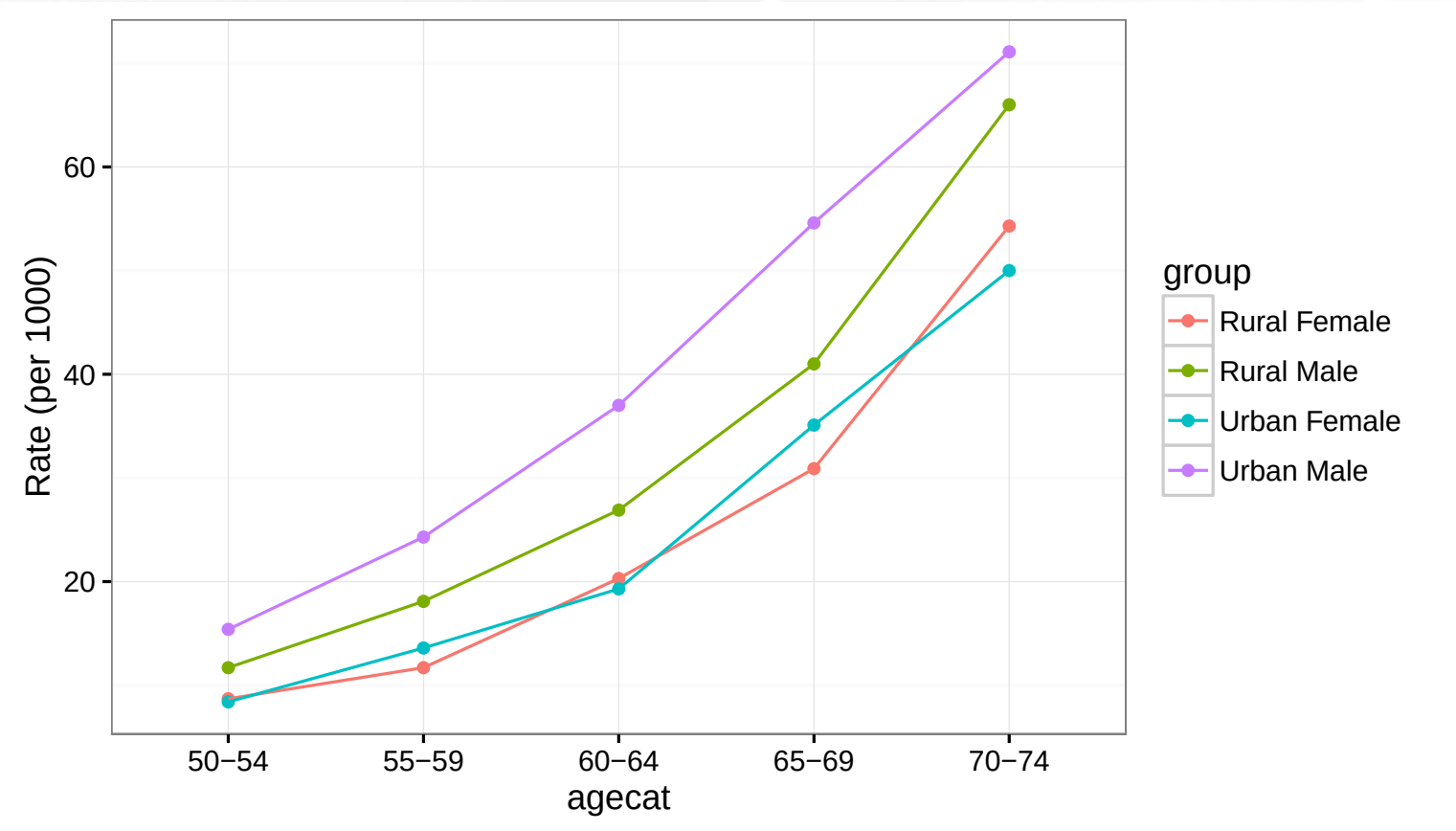


Dotplot versus barchart with facets





Grouped dotplot



Many categorical variables

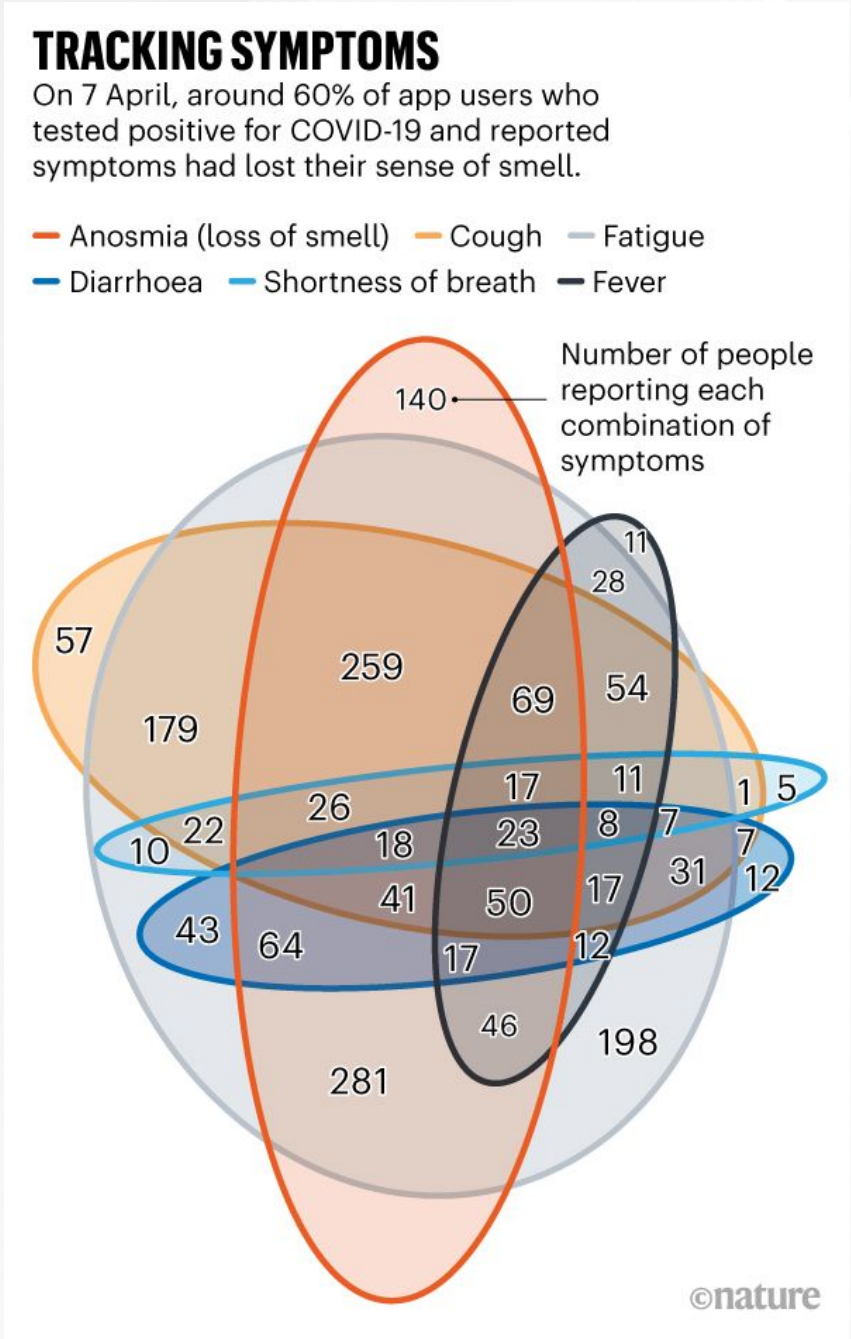
Example

Three different covid-19 symptoms; may occur together.
Eight different combinations

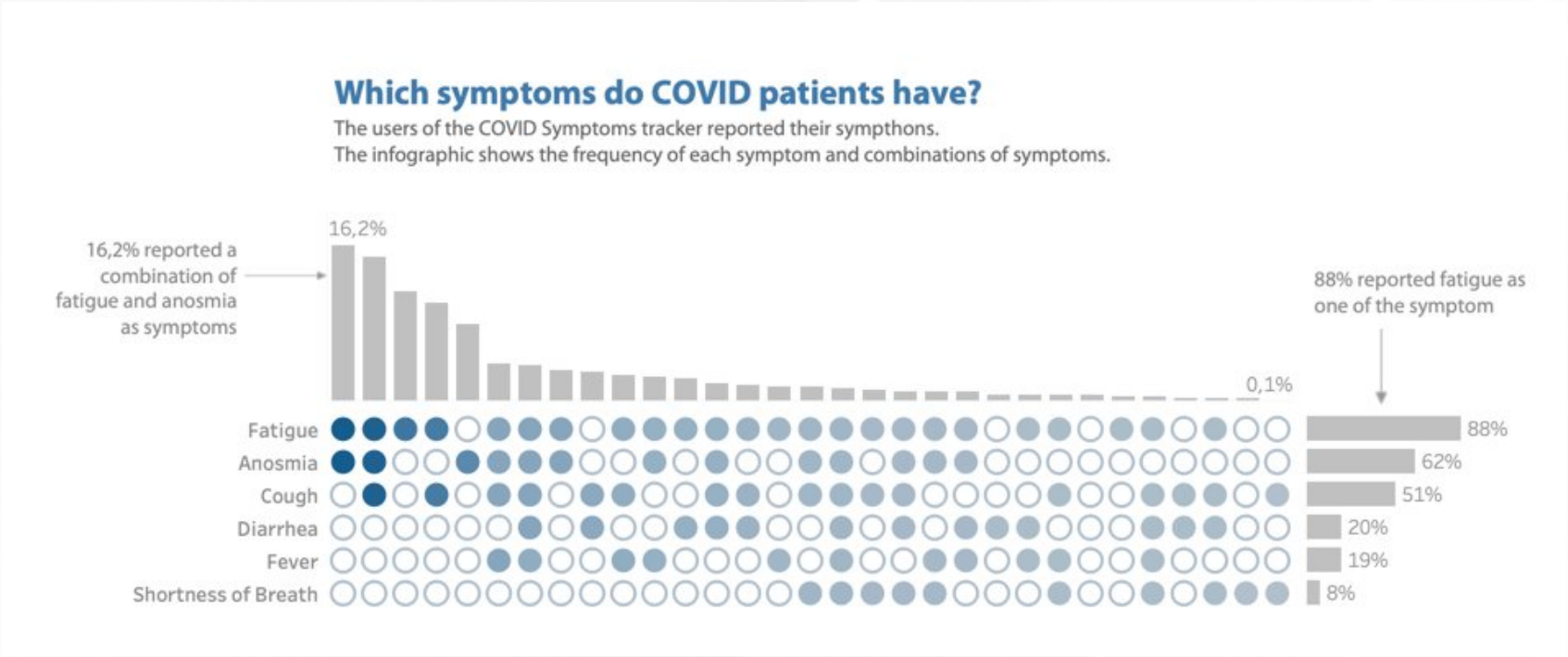
Anosmia	Fatigue	Cough	Nr of symptoms
No	No	No	0
Yes	No	No	1
No	Yes	No	1
No	No	Yes	1
Yes	Yes	No	2
Yes	No	Yes	2
No	Yes	Yes	2
Yes	Yes	Yes	3

With 6 symptoms there are 64 possible combinations

Covid-19 symptoms: Venn diagram



Covid-19 symptoms: upset plot



Outline

Exploratory Data Analysis (EDA)

Graphs: two or more categorical variables

Numeric variable by groups

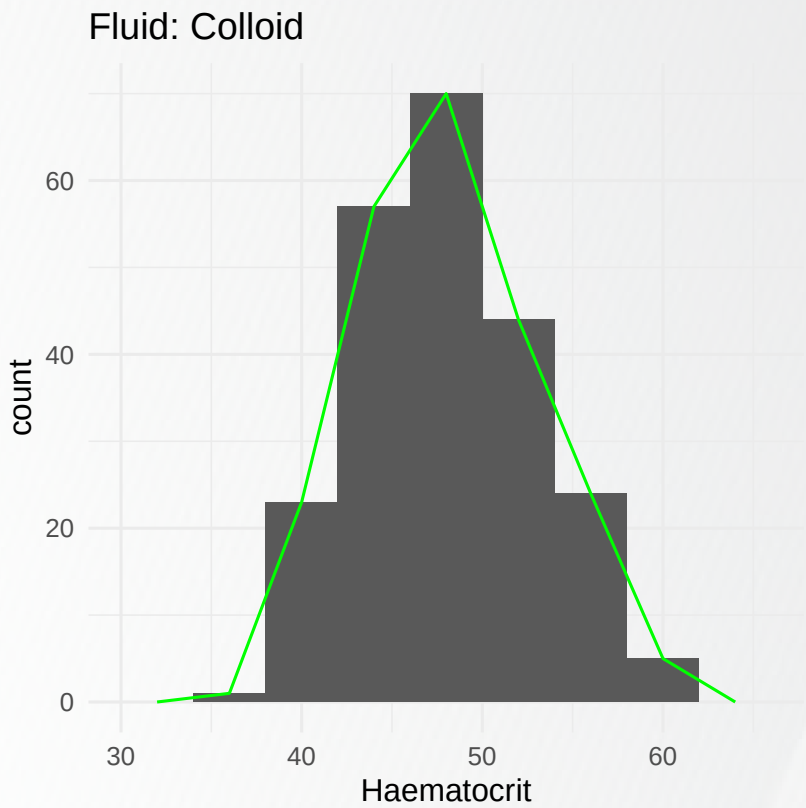
Two numeric variables

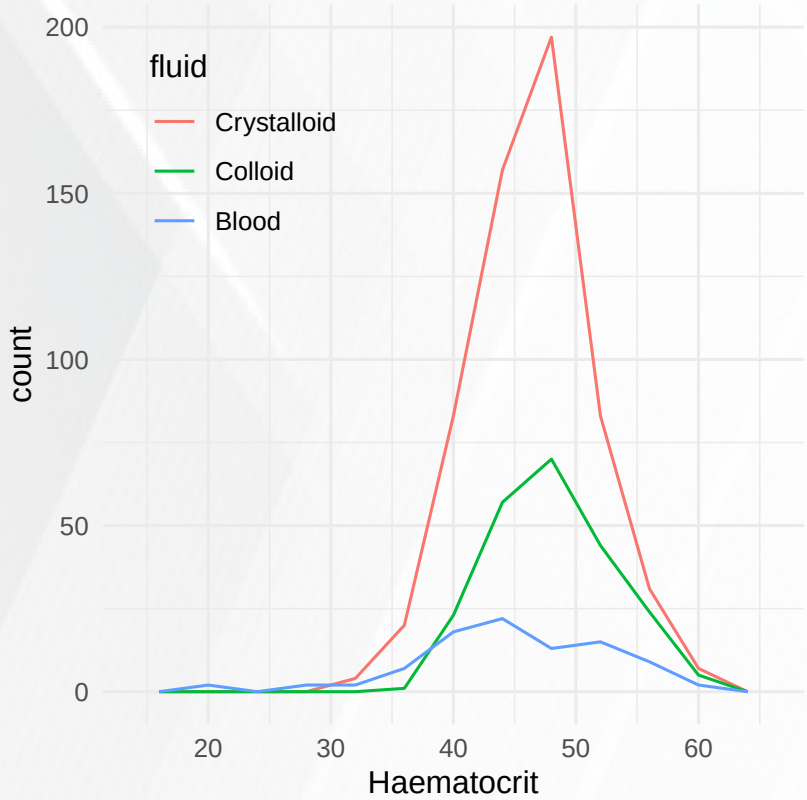
Data visualization in R

Single continuous variable by subgroup

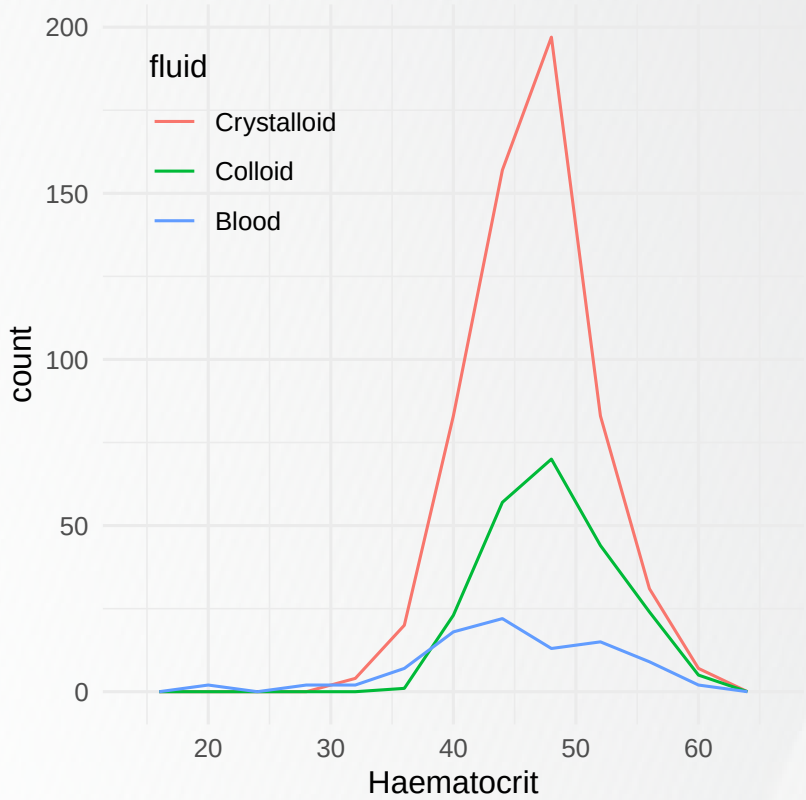
- Histogram, frequency polygon, density, ridgeplot

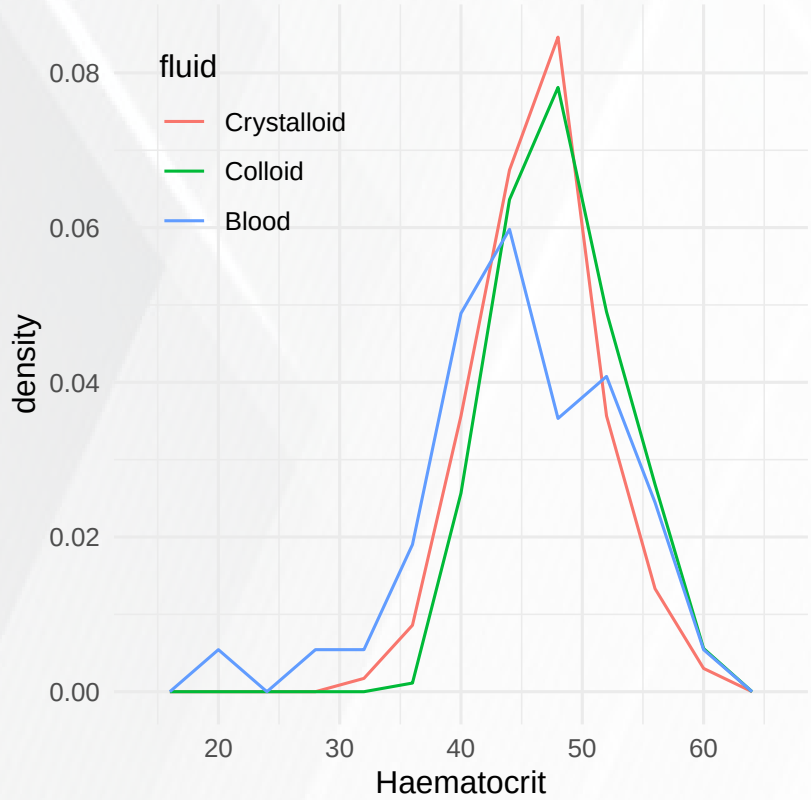
Histogram style for groups: “frequency polygon”



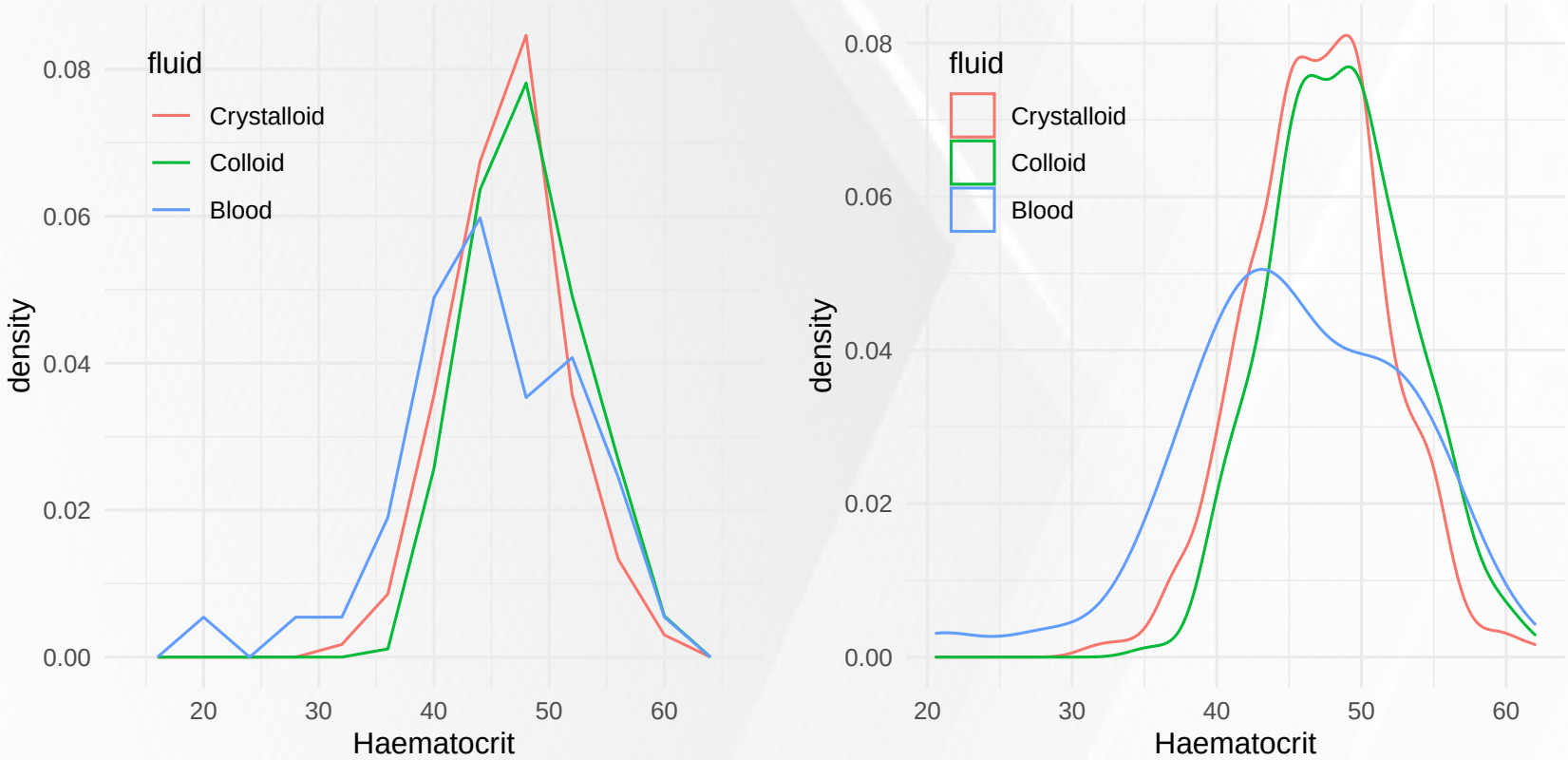


Comparison easier with standardized frequency polygon





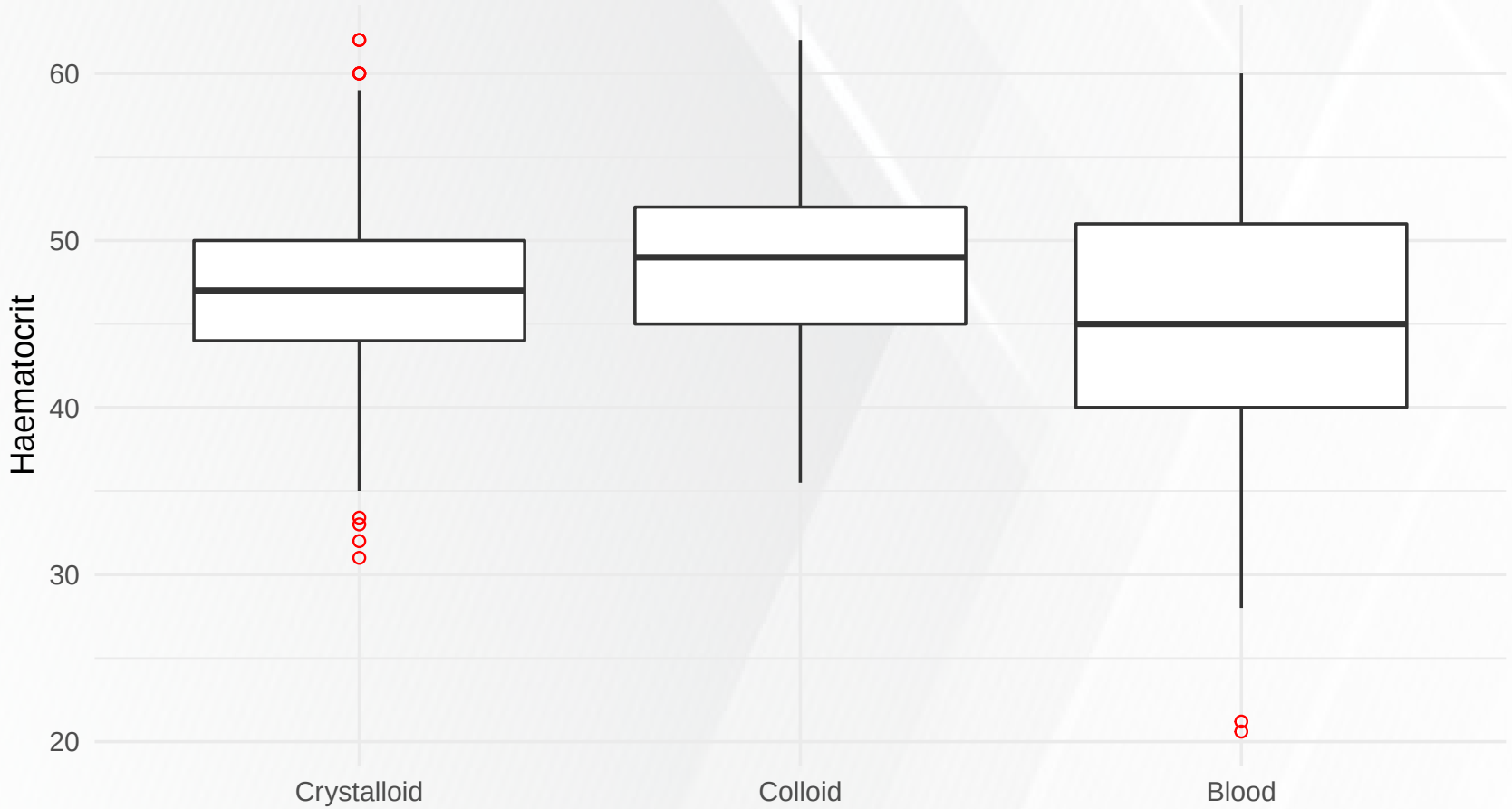
Smooth standardized frequency polygon: density plot



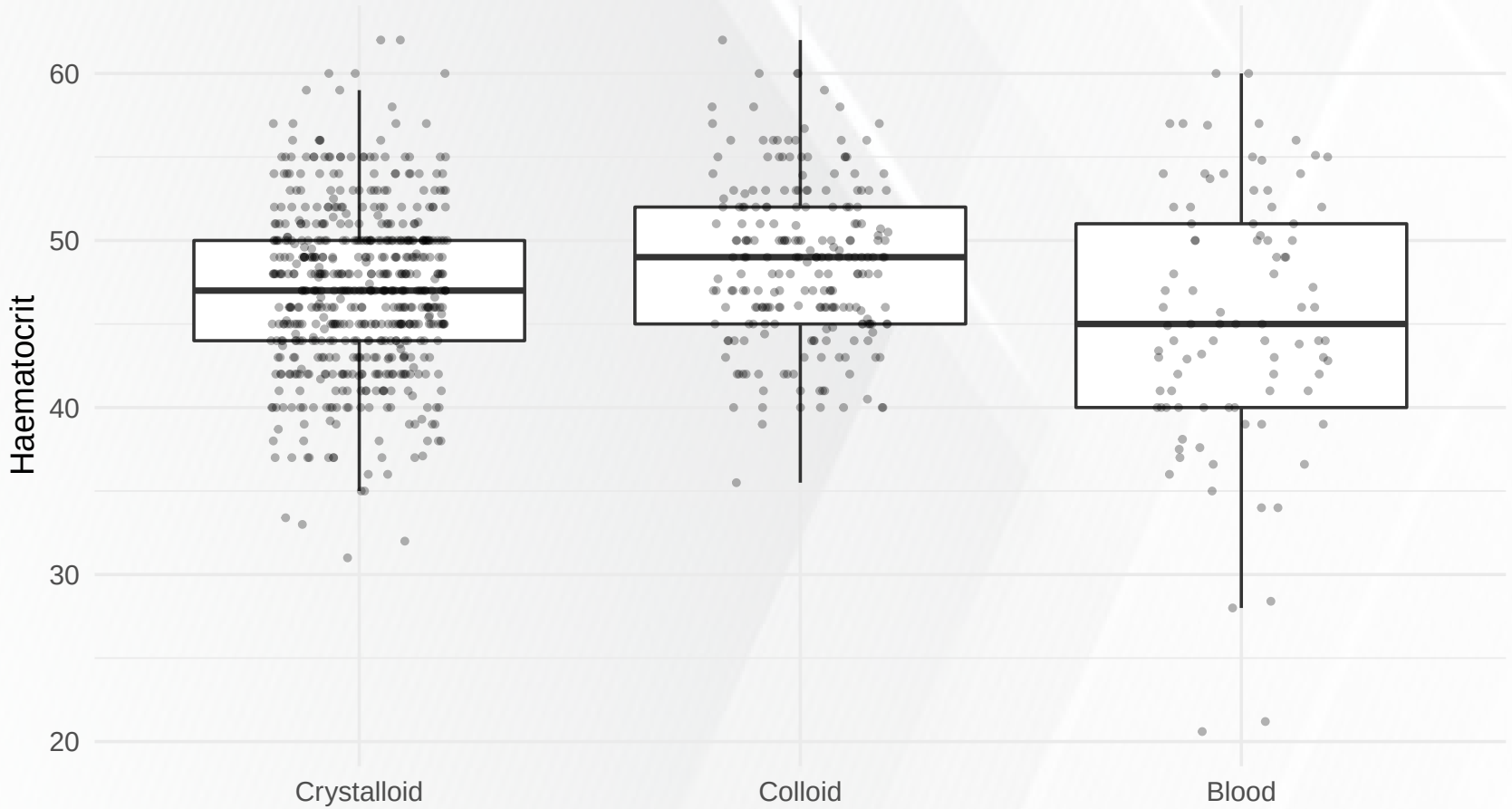
Single continuous variable by subgroup

- Histogram, frequency polygon, density, ridgeplot
- Boxplot, violin plot, raincloud plot

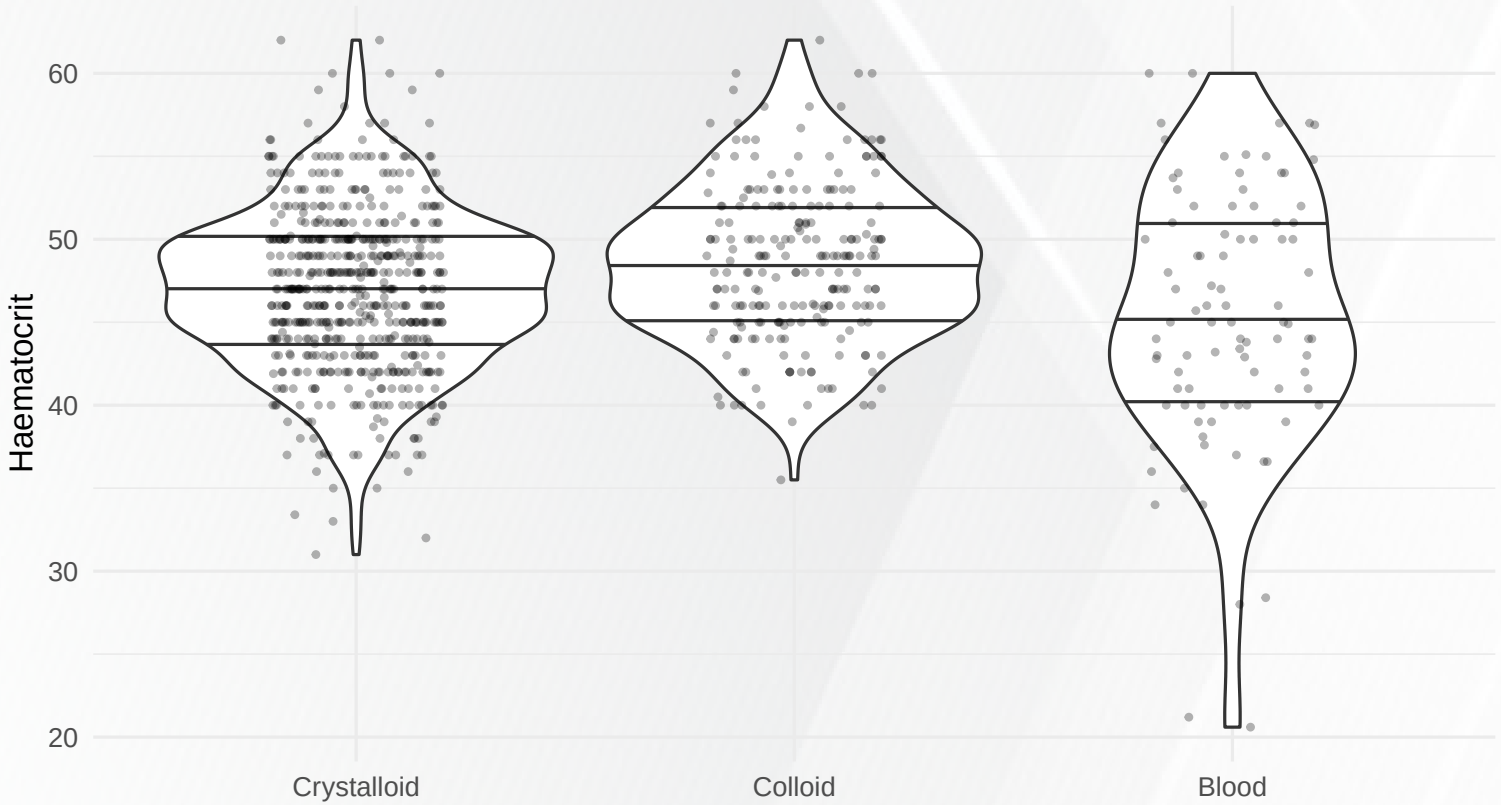
Boxplot for multiple groups



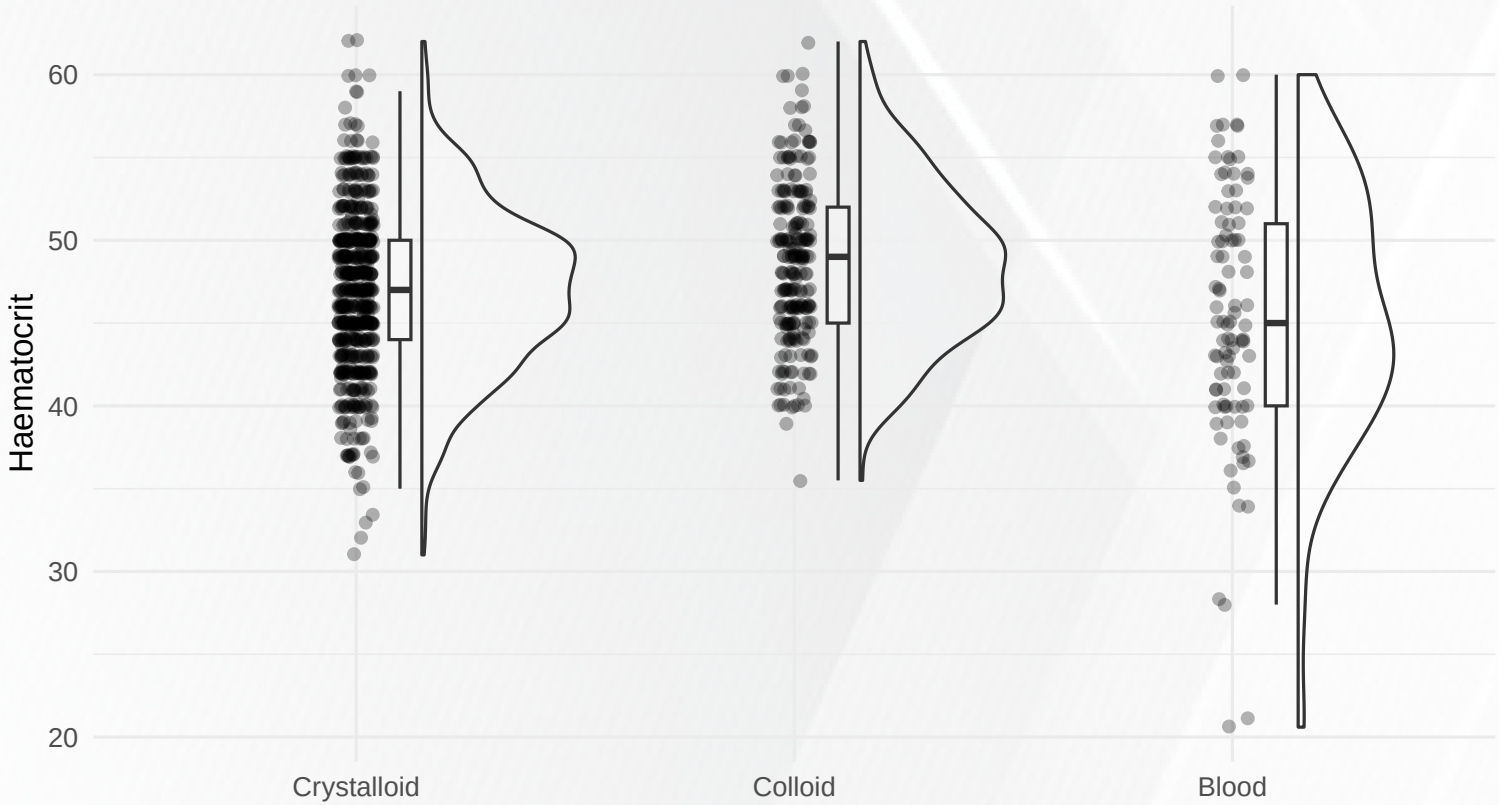
Boxplot, individual values added



Violin plot



Raincloud plot



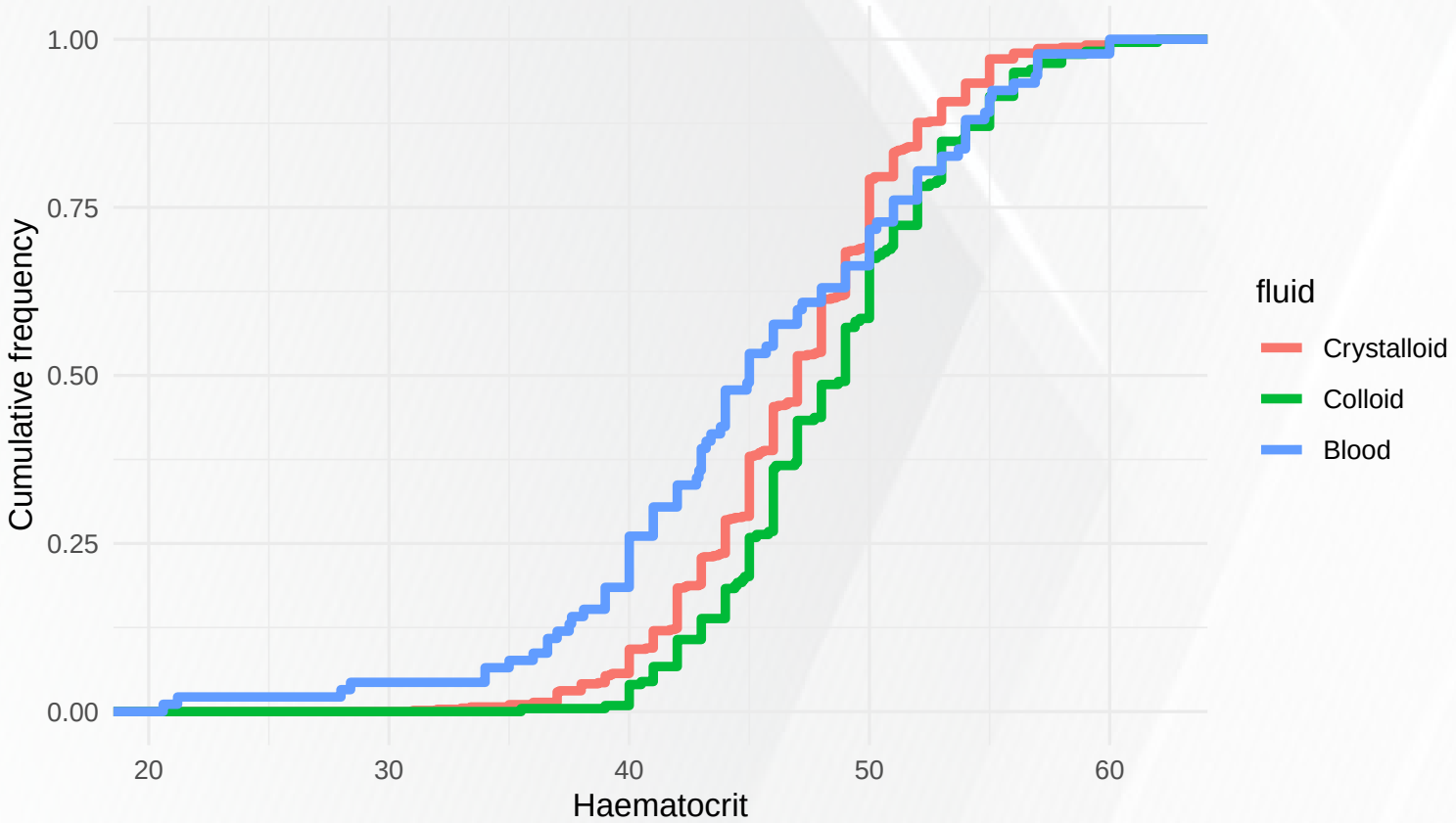
Single continuous variable by subgroup

- Histogram, frequency polygon, density, ridgeplot
- Boxplot, violin plot, raincloud plot
  Don't use dynamite plots

Single continuous variable by subgroup

- Histogram, frequency polygon, density, ridgeplot
- Boxplot, violin plot, raincloud plot
  Don't use dynamite plots
- Cumulative frequency ("empirical cumulative distribution function", `ecdf`)

ECDF plot


1.00
0.75
0.50
0.25
0.00
Cumulative frequency
20
30
40
50
60
Haematocrit
fluid
Crystalloid
Colloid
Blood

Outline

Exploratory Data Analysis (EDA)

Graphs: two or more categorical variables

Numeric variable by groups

Two numeric variables

Data visualization in R

Scatterplot; correlation coefficient ρ

- Strength of relation between numerical variables
  - standardized; not dependent on unit that is used

$$-1 \leq \rho \leq 1$$

  - positive correlation: if value of one variable increases, value of the other also tends to increase
  - negative correlation: if value of one variable increases, value of the other tends to decrease

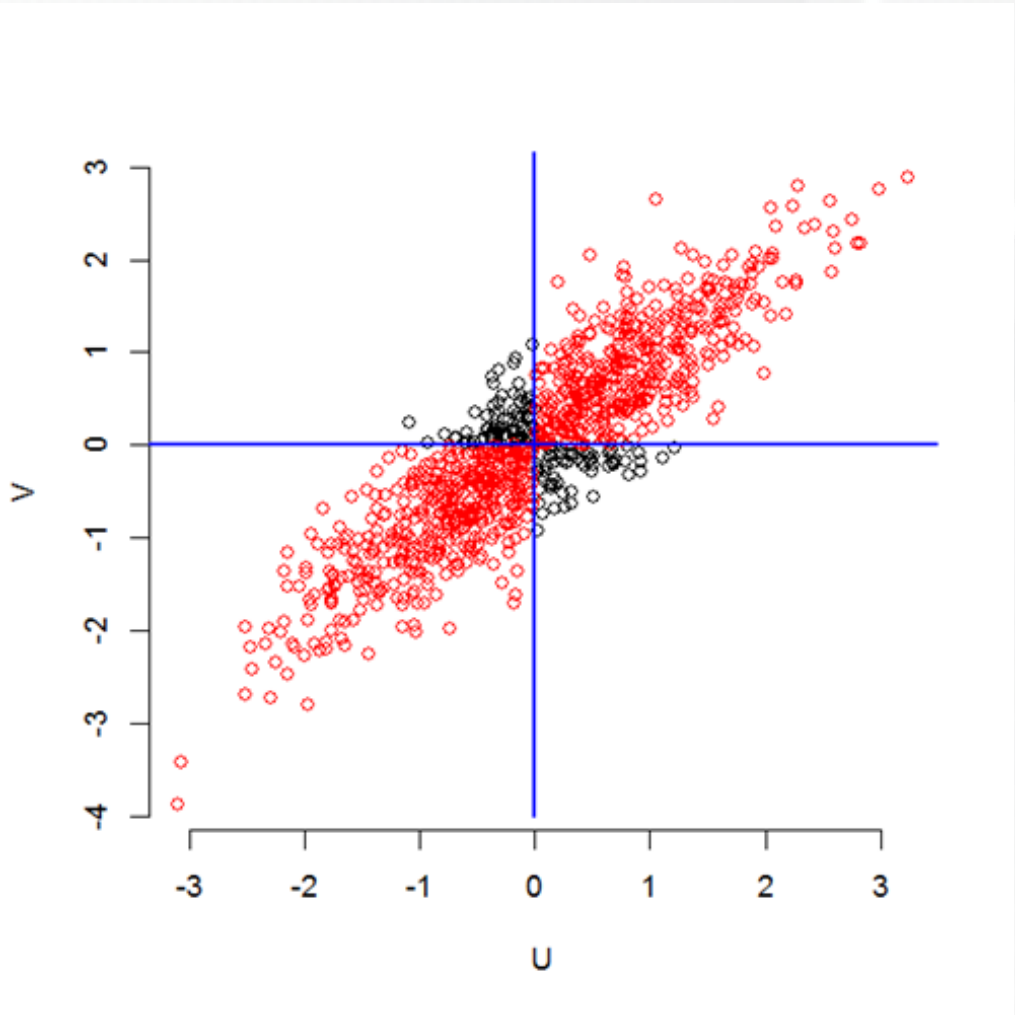
Scatterplot; correlation coefficient ρ

- Strength of relation between numerical variables
  - standardized; not dependent on unit that is used

$$-1 \leq \rho \leq 1$$

  - positive correlation: if value of one variable increases, value of the other also tends to increase
  - negative correlation: if value of one variable increases, value of the other tends to decrease
- Pearson's correlation coefficient

Calculation of Pearson's correlation



$$u_i = \frac{x_i - \overline{x}}{\mathrm{sd}_x}$$

$$v_i = \frac{y_i - \overline{y}}{\mathrm{sd}_y}$$

$$\rho = \frac{1}{N-1} \sum_{i=1}^{N} u_i \times v_i$$

N : sample size

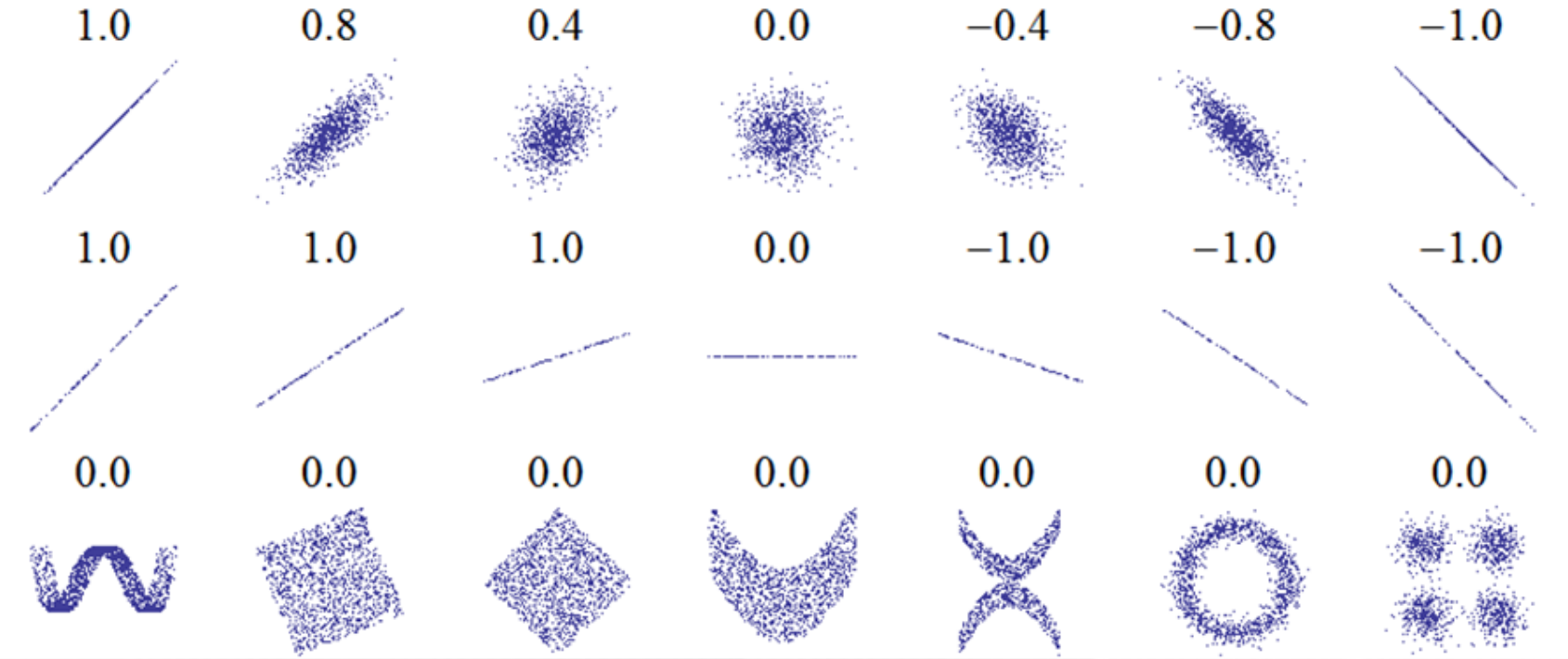
Scatterplot; correlation coefficient ρ

- Strength of relation between numerical variables
  - standardized; not dependent on unit that is used

$$-1 \leq \rho \leq 1$$

  - positive correlation: if value of one variable increases, value of the other also tends to increase
  - negative correlation: if value of one variable increases, value of the other tends to decrease
- Pearson's correlation coefficient
  - degree of linear relationship
  - $\rho = -1$ or $\rho = 1$: perfectly linear relationship
    All points on a straight line
  - $\rho = 0$: no linear relationship. Relationship can still be nonlinear

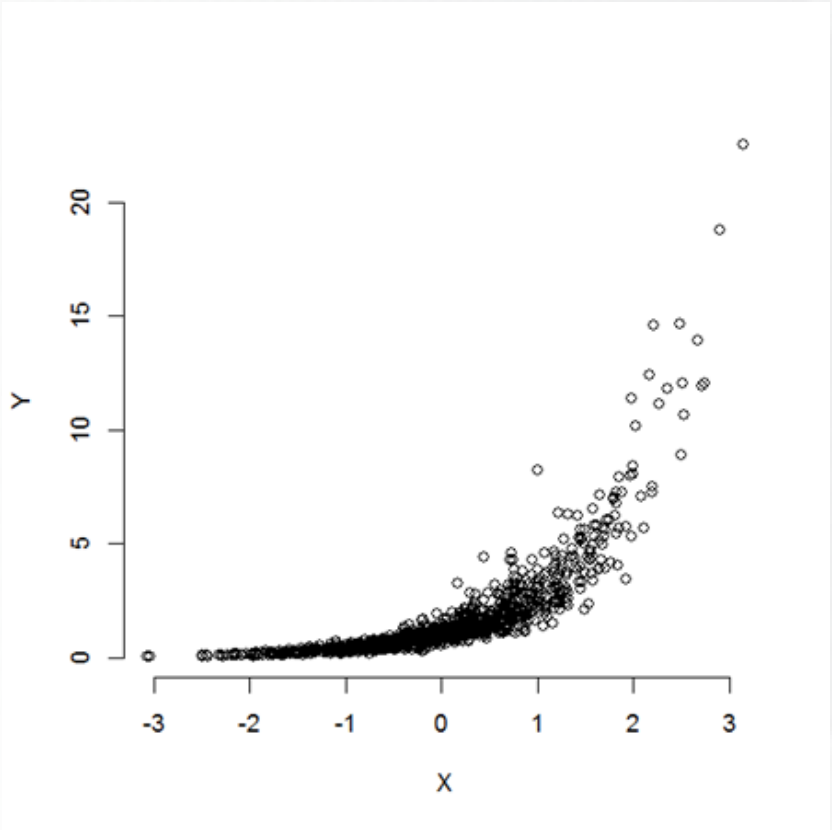
Pearson's correlation coefficient; examples

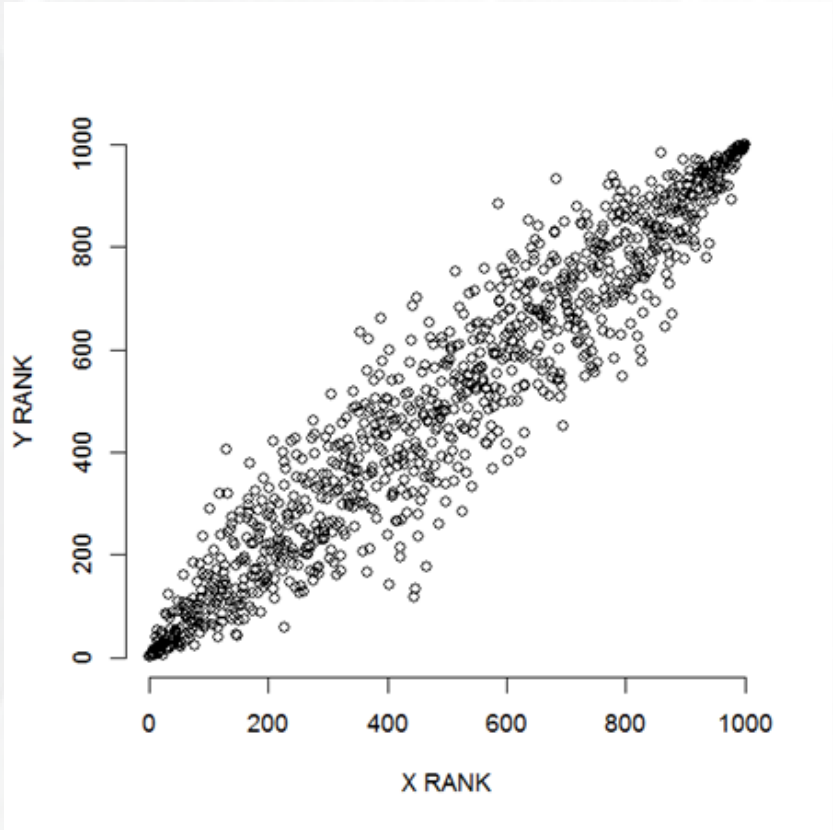


Spearman's rank correlation

- Calculation
  - assign to each value of variable x_i its rank:
    smallest value rank 1, second smallest rank 2 etc.
  - assign to each value of variable y_i its rank
  - compute Pearson's correlation between ranks
- Quantifies degree of monotone relationship
  - rank correlation +1:
    the relationship is positive and perfectly monotone
    (the larger x, the larger y)
  - rank correlation -1:
    the relationship is negative and perfectly monotone
    (the larger x, the smaller y)

Pearson's versus Spearman's; example





Pearson's correlation = 0.76

Spearman's correlation = 0.95

oucru

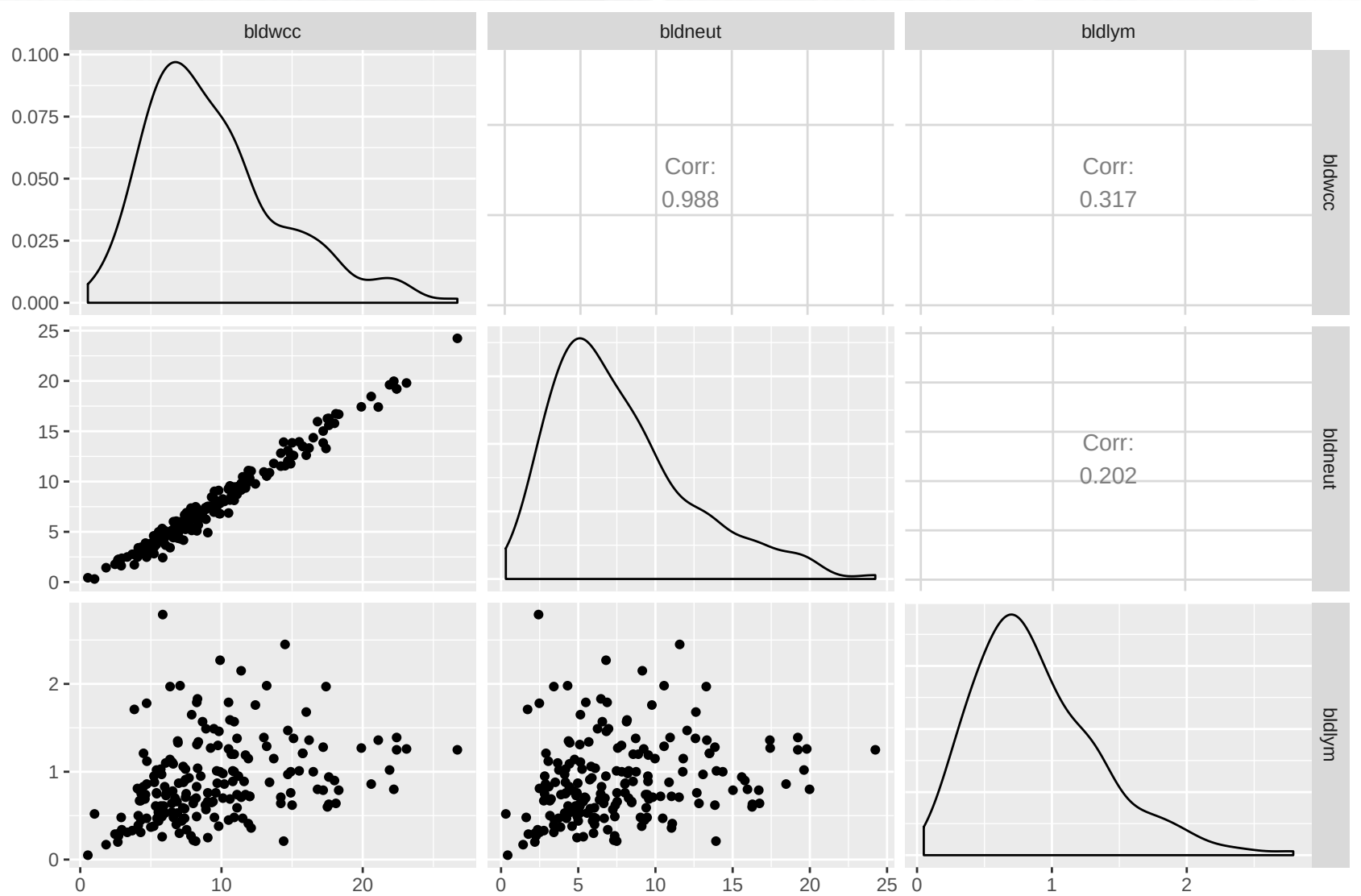
UNIVERSITY OF OXFORD

wellcome

Pearson's versus Spearman's; when?

- Pearson's correlation appropriate if both x and y have approximately symmetric distribution and the scatterplot shows a roughly elliptical pattern
- Otherwise
  - Try to make them symmeric/elliptic via transformation
  - Use Spearman's correlation

Pairwise comparison: scatterplot matrix



ChatGPT 5.2: “What are the best graphics for exploratory data analysis?”

Outline

Exploratory Data Analysis (EDA)

Graphs: two or more categorical variables

Numeric variable by groups

Two numeric variables

Data visualization in R

R graphics

- Base graphics: standard plotting commands
  `plot` most used
  others: `hist, boxplot, ...`

R graphics

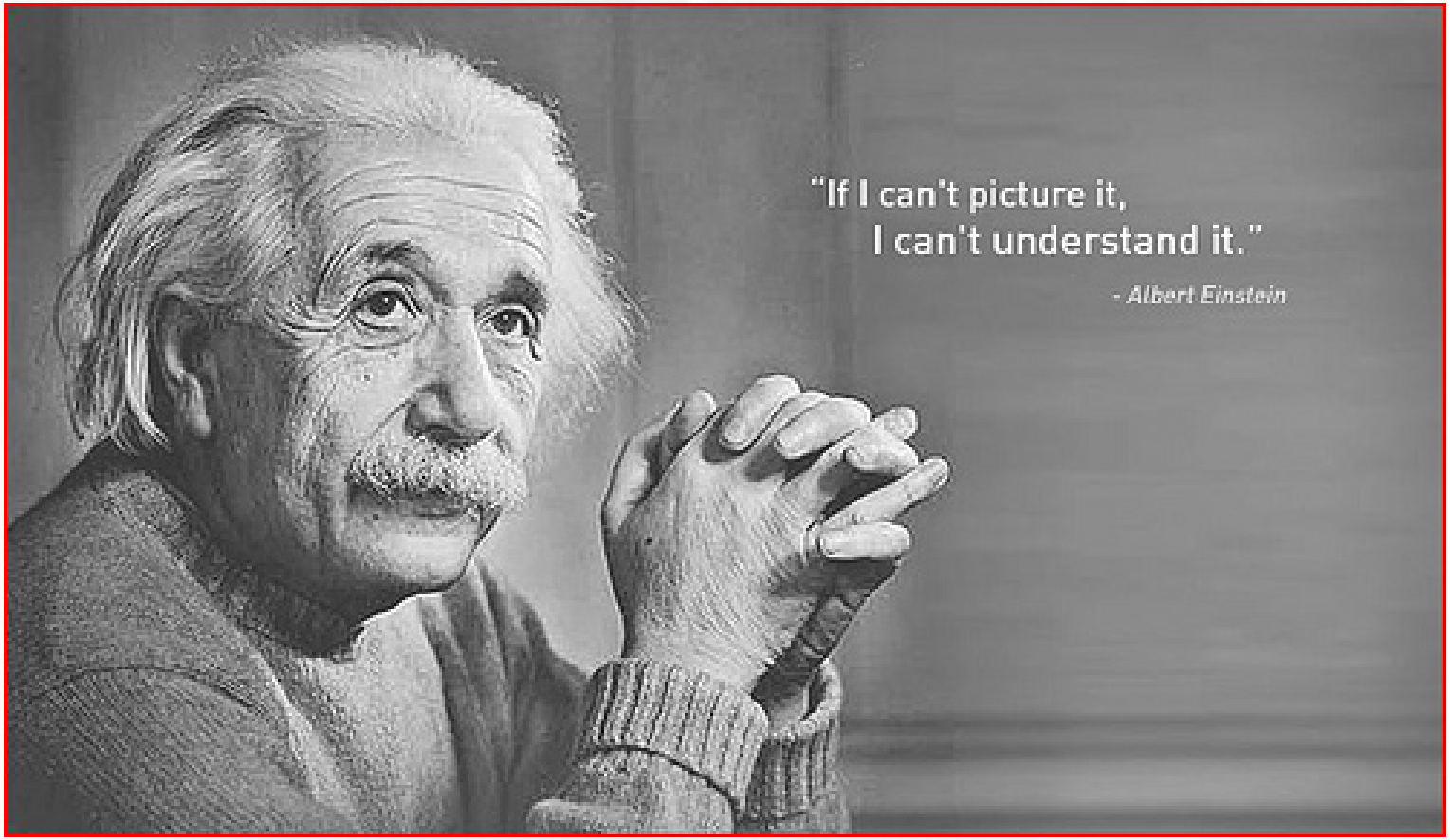
- Base graphics: standard plotting commands
  `plot` most used
  others: `hist, boxplot, ...`
- `ggplot2`: based on “the grammar of graphics”

R graphics

- Base graphics: standard plotting commands
  `plot` most used
  others: `hist, boxplot, ...`
- `ggplot2`: based on "the grammar of graphics"
- interactive plots (`rgl`, `plotly`)
- We mostly use `ggplot2` in this course

The ggplot2 package

- Website https://ggplot2.tidyverse.org documentation at https://ggplot2.tidyverse.org/reference
- Nice GUI (R package or Web application): https://dreamrs.github.io/esquisse
- Extensions: https://exts.ggplot2.tidyverse.org
- Books
  - ggplot2: Elegant Graphics for Data Analysis
  - Data Visualization
  - R Graphics Cookbook


"If I can't picture it,
I can't understand it."
- Albert Einstein