Data and Visualization Types

Nikodem Lewandowski

University of Gdańsk

Visualization types

- What visualization type should I choose?
  - It all depends on the data you have!
- There are two main types of data:
  - Categorical
  - Numeric

University of Gdańsk

Visualization types

- Many others to choose from, and an infinite number of combinations!

University of Gdańsk

Categorical Variables

- A categorical variable (also called qualitative variable) refers to a characteristic that can't be quantifiable.
- Categorical variables can be either:
  - Nominal, one that describes a name, label, or category without a natural order
  - Ordinal, its values are defined by an order relation between different categories

What's the difference?

- You **can't** compare and order nominal variables
- You **can** compare and order ordinal variables

University of Gdańsk

Nominal Variables

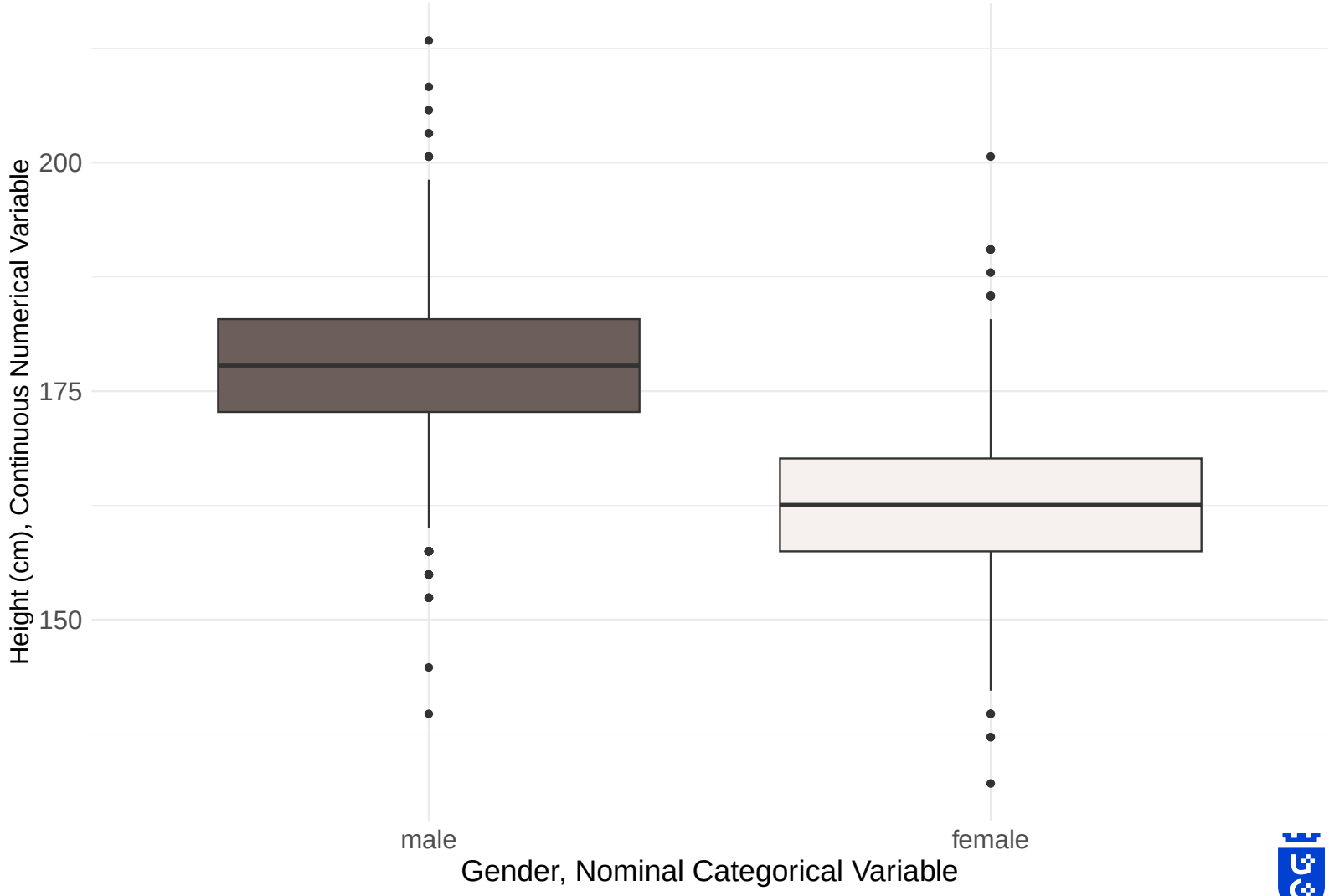
- A nominal variable is one that describes a name, label, or category without a natural order.
- Examples of nominal variables include:
  - Gender
  - Type of dwelling

What cookie do you like the most?

Cookies	Cookie.Lovers
Oreo	1000
Chocolate Cookie	1234
Oatmeal Cookie	2713

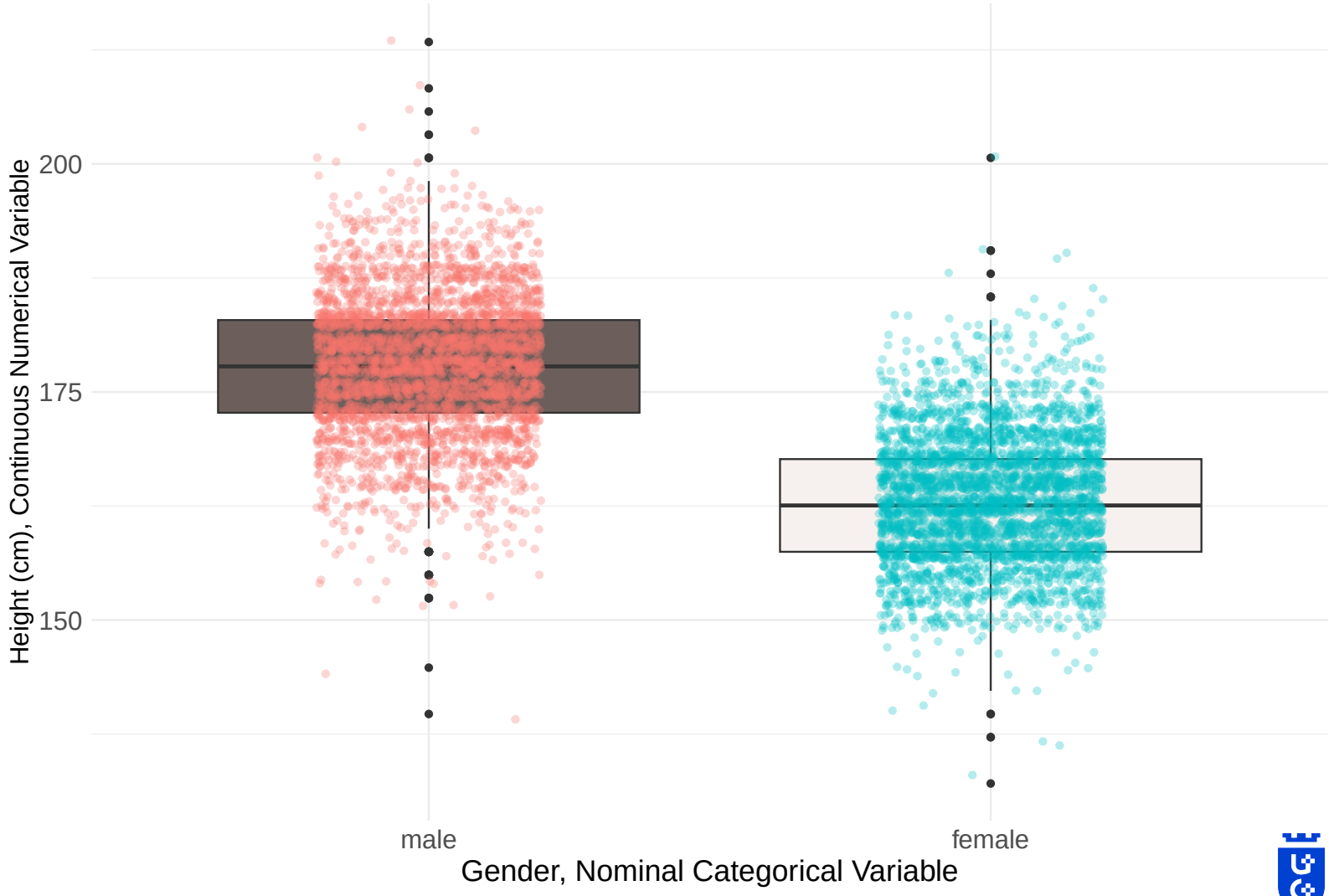
University of Gdańsk

Distribution of Heights by Gender



University of Gdańsk

Distribution of Heights by Gender with Jitter



University of Gdańsk

Code for the visualization above

```r
# packages and heights dataset is already loaded
heightsMODED <- heights
heightsMODED$height_cm <- heightsMODED$height * 2.54

custom_colors = c('#6C5F5B', '#F6F1EE') # hex colors (google for reference)

ggplot(heightsMODED, aes(x = sex, y = height_cm, fill = sex)) +
  geom_boxplot(custom_colors) +
  geom_jitter( # added on jitter plot
    aes(color = sex), width = 0.2, alpha = 0.3
    ) +
  labs(
    title = "",
    x = "Gender, Nominal Categorical Variable",
    y = "Height (cm), Continuous Numerical Variable"
  ) +
  theme_minimal()+
    theme(legend.position = "none",  # remove legend
        axis.text.x = element_text(size = 14), # change size of x axis text
        axis.text.y = element_text(size = 14),
        axis.title.x = element_text(size = 16), # change size of x axis title
        axis.title.y = element_text(size = 14)
        )
```

Ordinal Variables

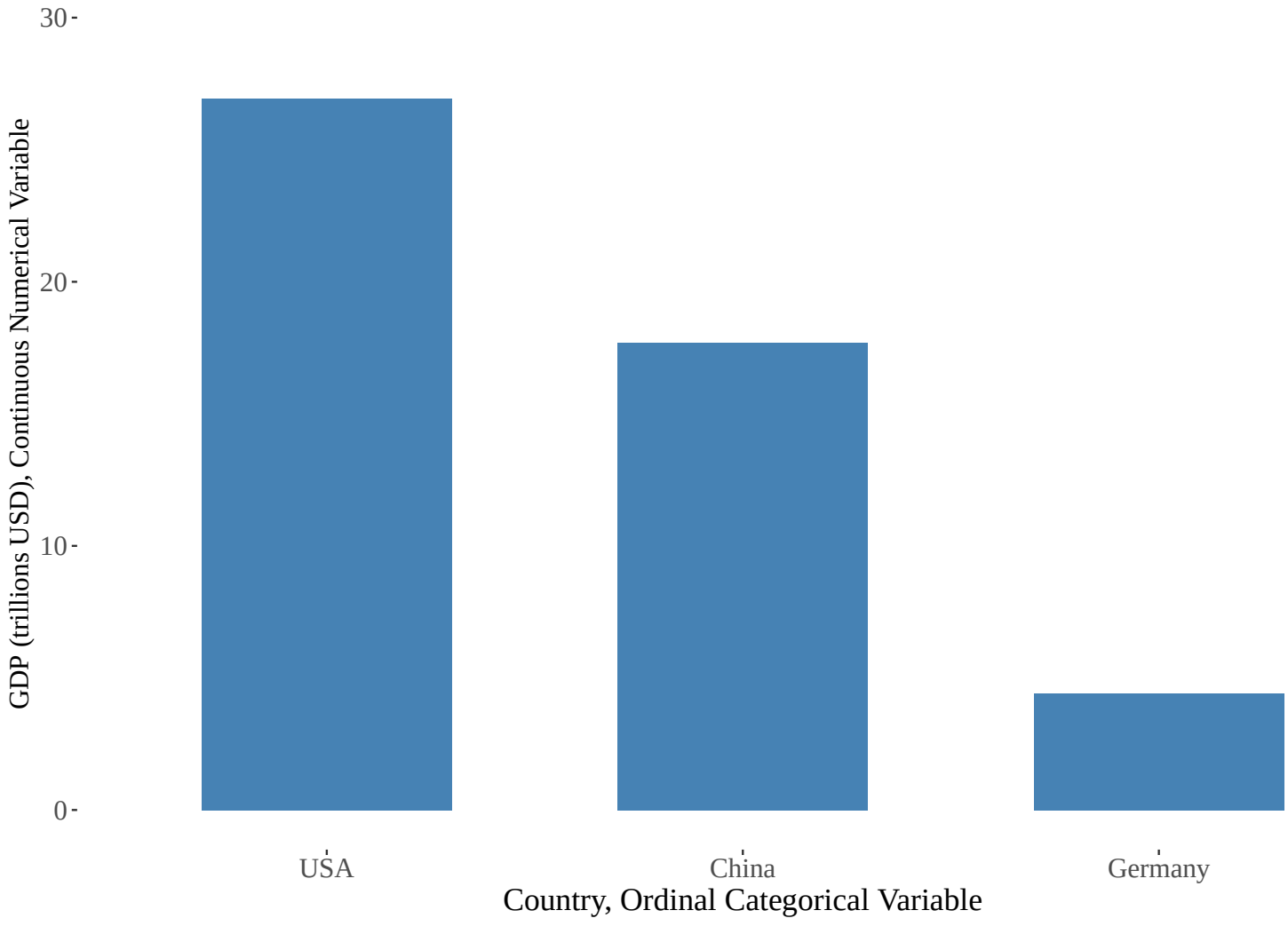
- There is some natural ordering, but the exact magnitude of the differences is (usually) not known.
- Examples of ordinal variables include:
  - Education level
  - Income level

Exam Results:

Category	Number.of.students
Excellent	5
Very good	12
Good	10
Bad	2

University of Gdańsk

GDP Estimation, 2023



University of Gdańsk

Code for the visualization above

```r
data <- data.frame(
  Country = c("USA", "China", "Germany"),
  GDP_2023 = c(26.94, 17.70, 4.42)
)

ggplot(data, aes(x = reorder(Country, -GDP_2023), y = GDP_2023)) +
  geom_bar(stat = "identity", fill = "steelblue", width= 0.6) +
  labs(
    title = "",
    x = "Country, Ordinal Categorical Variable",
    y = "GDP (trilions USD), Continuous Numerical Variable"
  ) +
  theme_tufte() +
  scale_y_continuous(limits= c(0, 30)) + # redefining y axis
  theme(axis.text.x = element_text(size = 14),
        axis.text.y = element_text(size = 14),
        axis.title.x = element_text(size = 16),
        axis.title.y = element_text(size = 14)
        )
```

University of Gdańsk

Numeric Variables

- A numeric variable (also called quantitative variable) is a quantifiable characteristic whose values are numbers.
- Numeric variables may be either continuous or discrete.
- A category might be represented by a single number or by a distribution of numbers.

University of Gdańsk

Continuous Variables

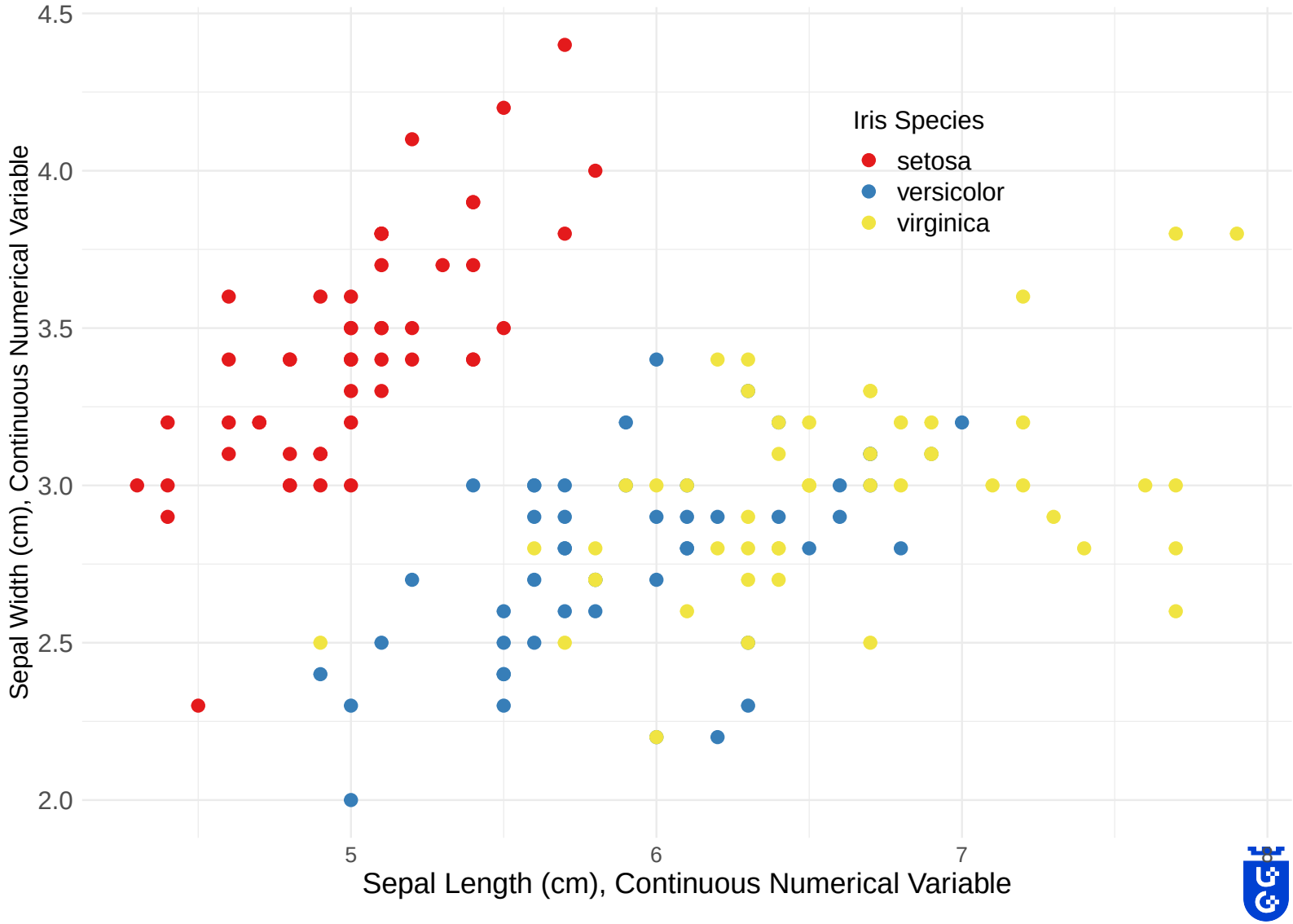
- A variable is said to be continuous if it can assume an infinite number of real values within a given interval.
- Examples of continuous variables include:
  - Height
  - Temperature

Example Table:

Time.hours	Temperature
0	23.3
1	25.7
2	27.1

University of Gdańsk

Iris Sepal Length vs. Sepal Width



University of Gdańsk

Iris sepal

University of Gdańsk

Code for the visualization above

```r
custom_colors <- c("#E41A1C", "#377EB8", "#f0e442")

ggplot(iris, aes(x = Sepal.Length, y = Sepal.Width, color = Species)) +
  geom_point(size = 3) +
  labs(
    x = "Sepal Length (cm), Continuous Numerical Variable",
    y = "Sepal Width (cm), Continuous Numerical Variable",
    color = 'Iris Species'  # Custom legend title
  ) +
  theme_minimal() +
   scale_color_manual(values = custom_colors) +
  theme(
    legend.position = c(0.7, 0.8),  # Custom legend position
    axis.text.x = element_text(size = 12),
    axis.text.y = element_text(size = 14),
    axis.title.x = element_text(size = 16),
    axis.title.y = element_text(size = 14),
    legend.text = element_text(size = 14),
    legend.title = element_text(size = 14)
  )
```

University of Gdańsk

Discrete Variables

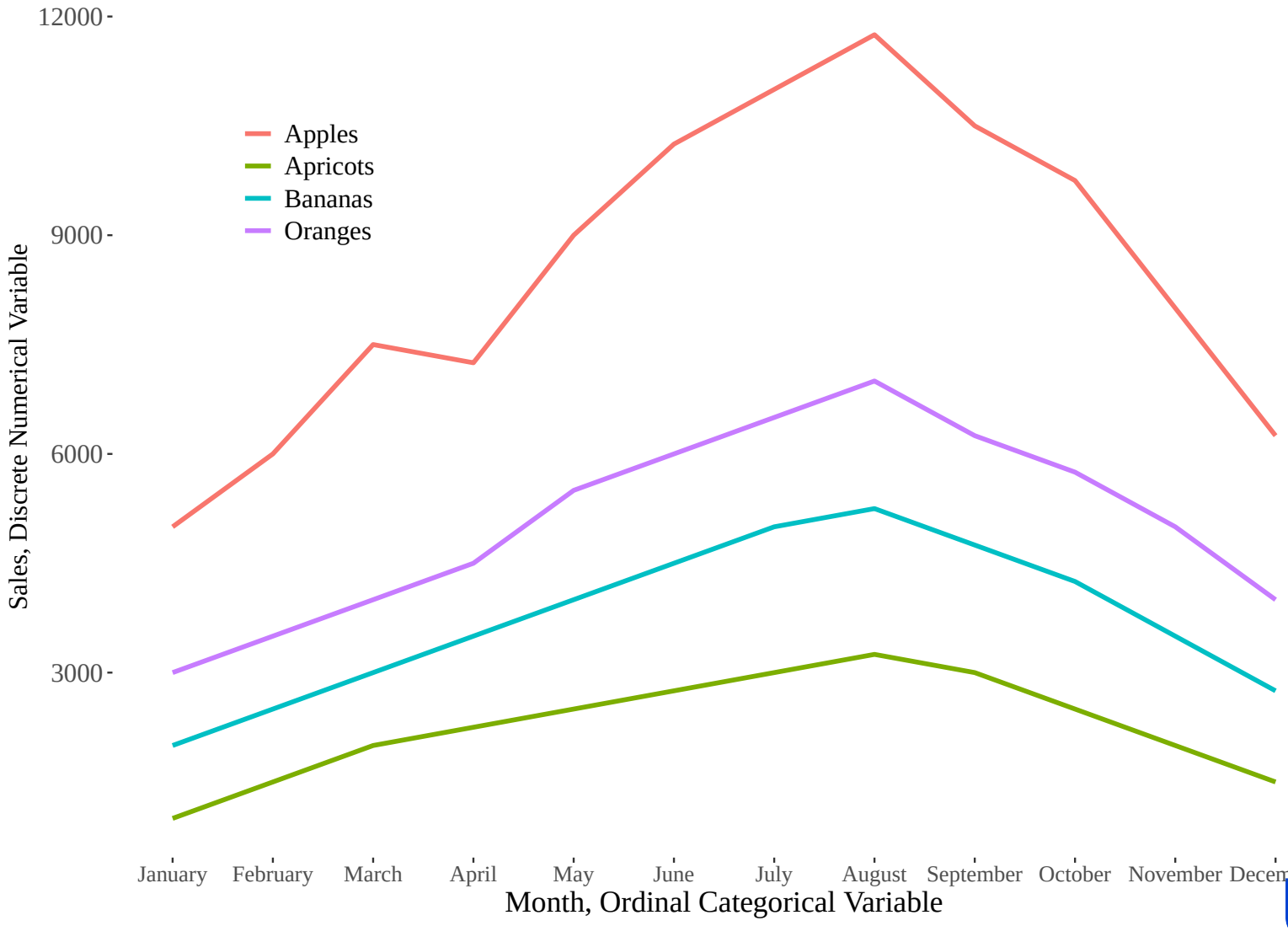
- A discrete variable can assume only a finite number of real values within a given interval.
- Examples of discrete variables include:
  - Number of customers
  - Number of products sold

Example Table:

Items	Sells
Bananas	10
Apples	15
Oranges	20

University of Gdańsk

Monthly Sales of Fruits, 2022



University of Gdańsk

Code for the visualization above

```r
sales_data <- data.frame(
    Month = c("January", "February", "March", "April", "May", "June",
     "July", "August", "September", "October", "November", "December"),
    Apples = c(5000, 6000, 7500, 7250, 9000, 10250, 11000, 11750, 10500, 9750, 8000, 6250),
    Oranges = c(3000, 3500, 4000, 4500, 5500, 6000, 6500, 7000, 6250, 5750, 5000, 4000),
    Bananas = c(2000, 2500, 3000, 3500, 4000, 4500, 5000, 5250, 4750, 4250, 3500, 2750),
    Apricots = c(1000, 1500, 2000, 2250, 2500, 2750, 3000, 3250, 3000, 2500, 2000, 1500))

melted_data <- sales_data %>%
  gather(Fruits, Sales, -Month) # gather to convert to long format

melted_data$Month <- factor(melted_data$Month, levels = c("January", "February", "March",
 "April", "May", "June", "July", "August", "September", "October", "November", "December"))
melted_data$Fruits <- as.factor(melted_data$Fruits) # converting strings to factors


ggplot(melted_data, aes(x = Month, y = Sales, color = Fruits)) +
  geom_line(aes(group = Fruits), linewidth= 1.2) +  # group to specify the variable to group by
  labs(x = "Month, Ordinal Categorical Variable",
       y = "Sales, Discrete Numerical Variable",
       color = "")+ # remove legend title
    theme_tufte() +
    theme(legend.position = c(0.16, 0.8), # custom legend position
        axis.text.x = element_text(size = 12),
        axis.text.y = element_text(size = 14),
        axis.title.x = element_text(size = 16),
        axis.title.y = element_text(size = 14),
        legend.text = element_text(size = 14)
        )
```

of Gdańsk

Additional Resources

Cool youtube videos to checkout:

- Science of Data Visualization | Bar, scatter plot, line...
- How To Choose The Right Graph

University of Gdańsk