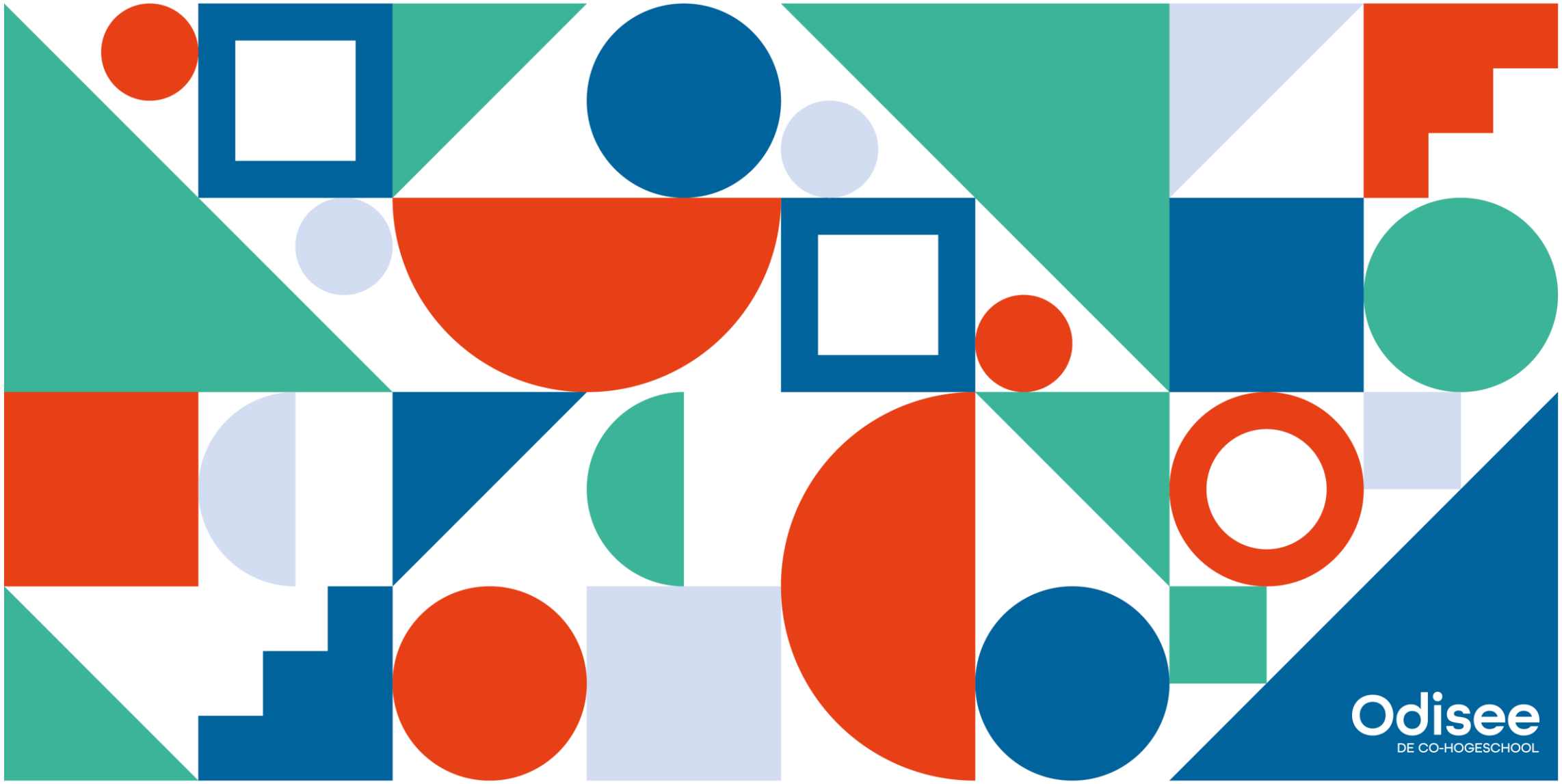
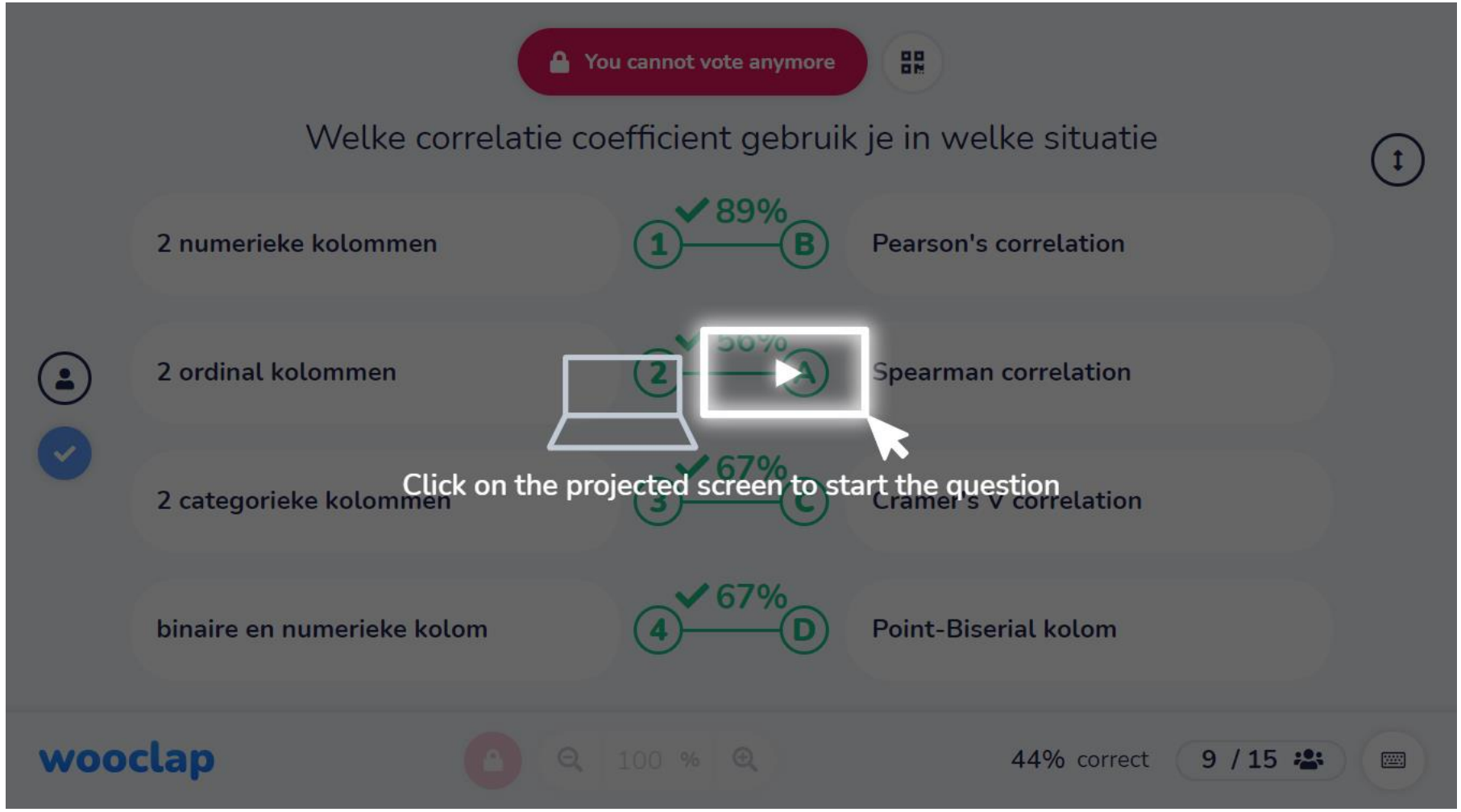
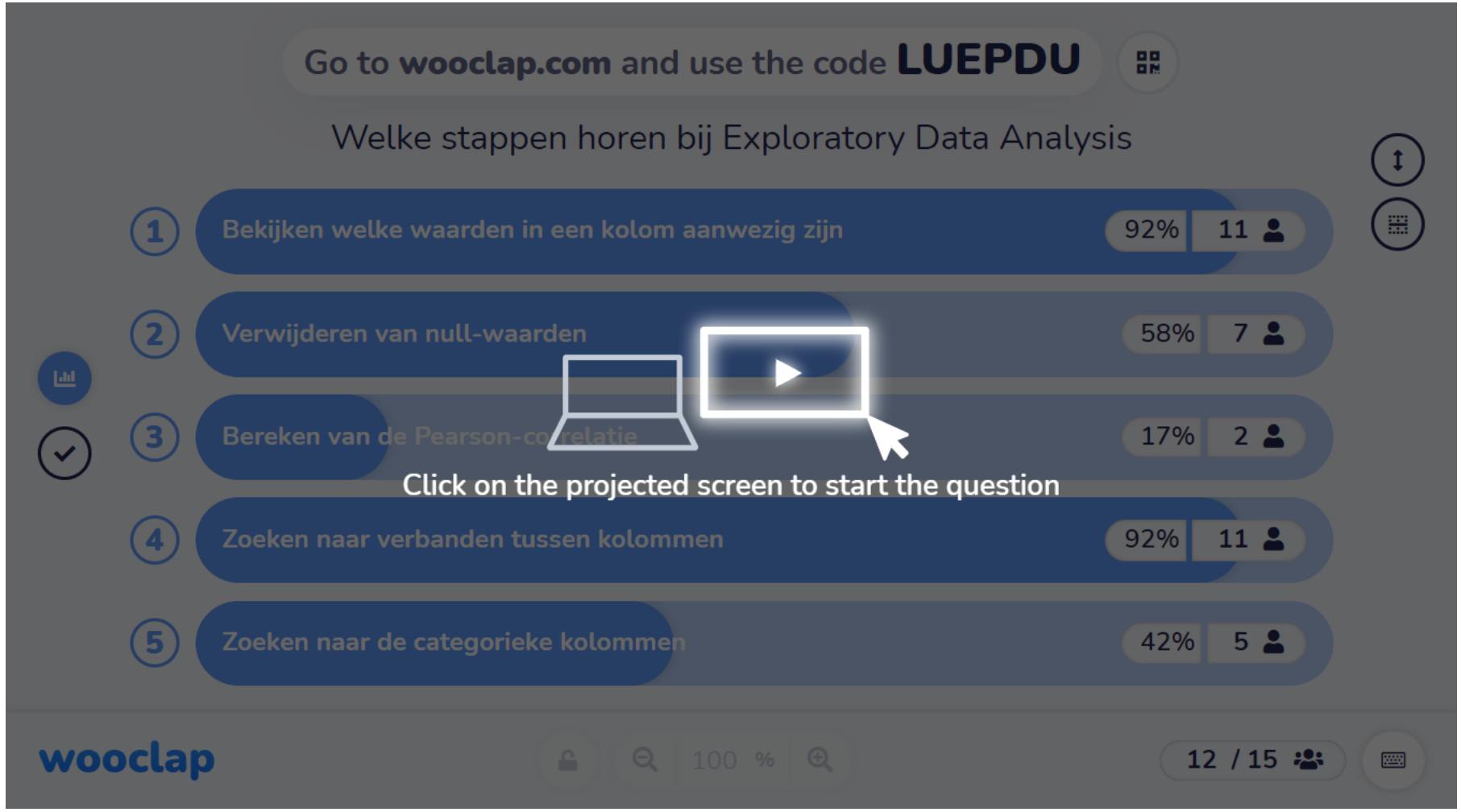

Odisee
DE CO-HOGESCHOOL

# Data Science – week 5

Jens Baetens


How to participate?
Click on the projected screen to start the question
Copy participation link
wooclap
100 %
15


You cannot vote anymore
Welke correlatie coefficient gebruik je in welke situatie
2 numerieke kolommen
89%
1
B
Pearson's correlation
2 ordinal kolommen
2
A
Spearman correlation
2 categorieke kolommen
67%
3
C
Cramer's V correlation
binaire en numerieke kolom
67%
4
D
Point-Biserial kolom
Click on the projected screen to start the question
wooclap
100 %
44% correct
9 / 15


Go to wooclap.com and use the code LUEPDU
Welke stappen horen bij Exploratory Data Analysis
1 Bekijken welke waarden in een kolom aanwezig zijn 92% 11
2 Verwijderen van null-waarden 58% 7
3 Bereken van de Pearson-correlatie 17% 2
Click on the projected screen to start the question
4 Zoeken naar verbanden tussen kolommen 92% 11
5 Zoeken naar de categorieke kolommen 42% 5
wooclap
100 %
12 / 15

Go to **wooclap.com** and use the code **LUEPDU**

Welke methode kan gebruikt worden voor outliers te detecteren?

| | | | |
|---|---|---|---|
| 1 | Interkwartielafstand methode | 83% | 10 |
| 2 | Point Biserial | 0% | 0 |
| 3 | Datatypes detecteren | 0% | 0 |
| 4 | Standaardafwijking berekenen | 17% | 2 |
| 5 | Data Masking | 0% | 0 |

wooclap

100 %

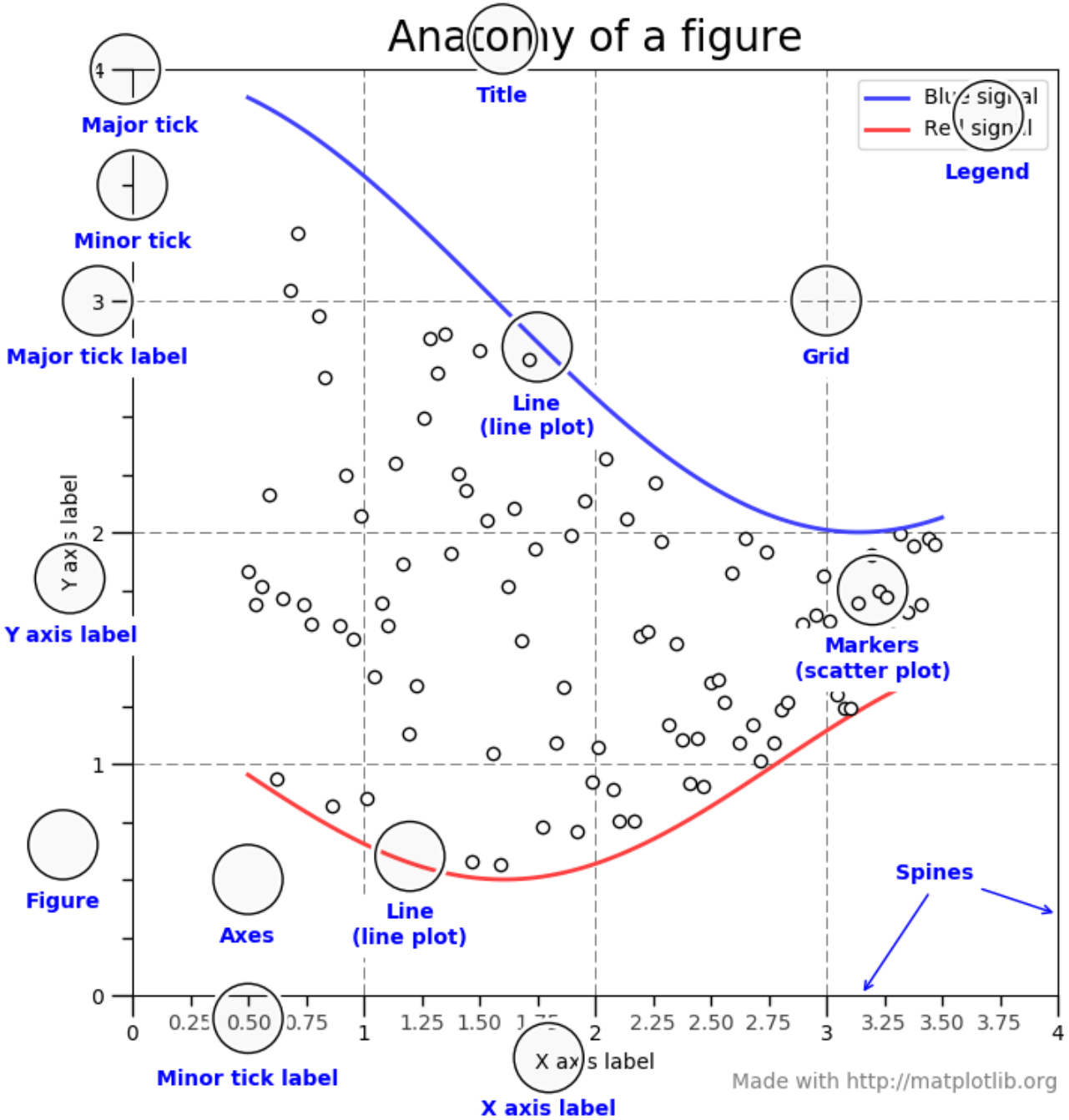
12 / 15

# Data visualisation

# Hoe visualiseer je data grafisch met python

- 3 veelgebruikte manieren
  - Matplotlib
  - .plot() functie in pandas
  - Seaborn

# Matpotlib

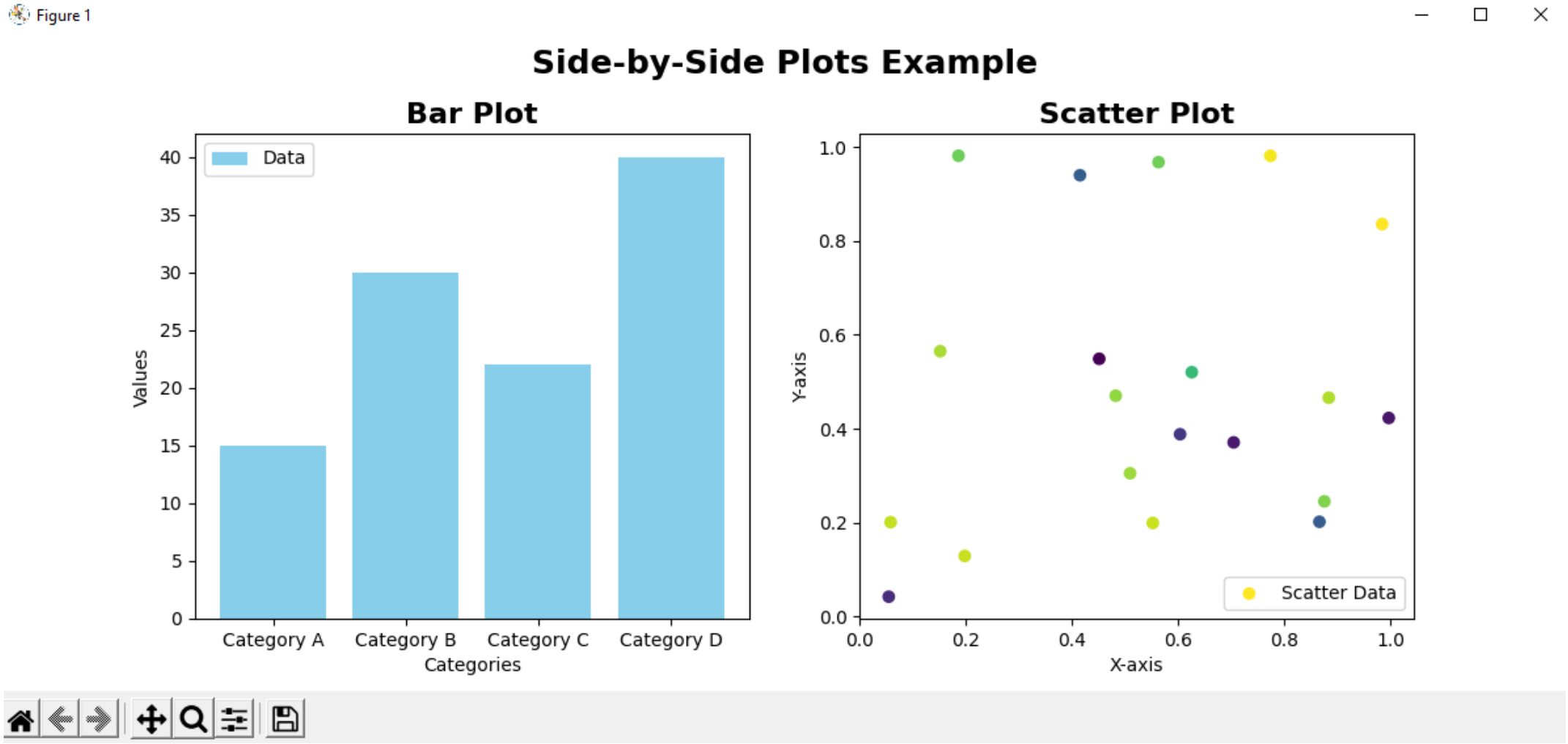

Anatomy of a figure
Title
Major tick
Minor tick
Major tick label
Legend
Grid
Line
(line plot)
Y axis label
Markers
(scatter plot)
Figure
Axes
Line
(line plot)
Spines
Minor tick label
X axis label
Made with http://matplotlib.org

# Matplotlib example

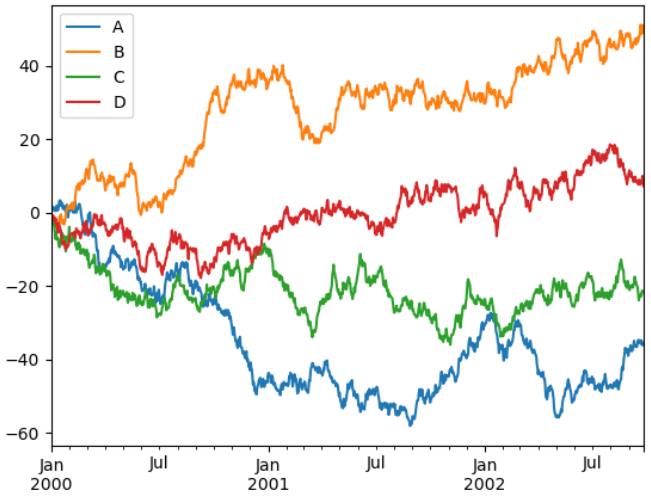
```
import matplotlib.pyplot as plt
import numpy as np

# Sample data
categories = ['Category A', 'Category B', 'Category C', 'Category D']
bar_data = [15, 30, 22, 40]

x_scatter = np.random.rand(20)
y_scatter = np.random.rand(20)
scatter_colors = np.random.rand(20)
scatter_markers = ['o', 's', 'D', '^', 'v', '<', '>', 'p', '*', 'h', 'H',

# Create a figure with two subplots
fig, (ax1, ax2) = plt.subplots(1, 2, figsize=(12, 5))

# Bar plot
ax1.bar(categories, bar_data, color='skyblue', label='Data')
ax1.set_title("Bar Plot", fontsize=16, fontweight='bold')
ax1.set_xlabel("Categories")
ax1.set_ylabel("Values")
ax1.legend()
```

```
# Scatter plot
sc = ax2.scatter(x_scatter, y_scatter, c=scatter_colors, marker=scatter_mar
ax2.set_title("Scatter Plot", fontsize=16, fontweight='bold')
ax2.set_xlabel("X-axis")
ax2.set_ylabel("Y-axis")
ax2.legend()

# Adding a subtitle using text annotations
fig.suptitle("Side-by-Side Plots Example", fontsize=18, fontweight='bold')
fig.subplots_adjust(top=0.85)  # Adjust the spacing for the title and subti

# Adding ticks to the scatter plot
ax2.set_xticks(np.arange(0, 1.1, 0.2))
ax2.set_yticks(np.arange(0, 1.1, 0.2))

plt.show()
```

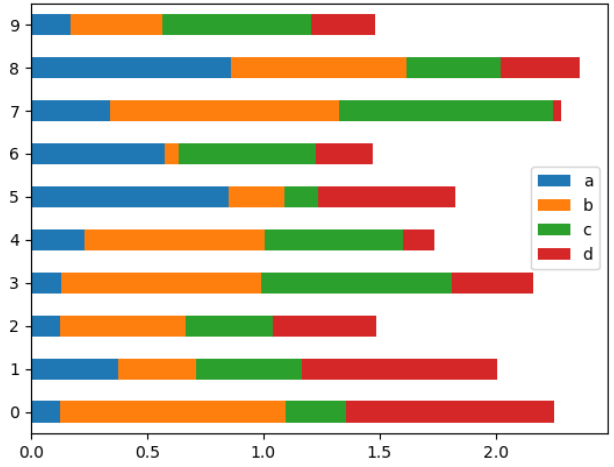
# Matplotlib example



# Pandas .plot()




```
df = pd.DataFrame(np.random.randn(1000, 4), index=ts.index, columns=list("ABCD"))

df = df.cumsum()

plt.figure();

df.plot();
```




```
df2.plot.barh(stacked=True);
```

# Pandas example

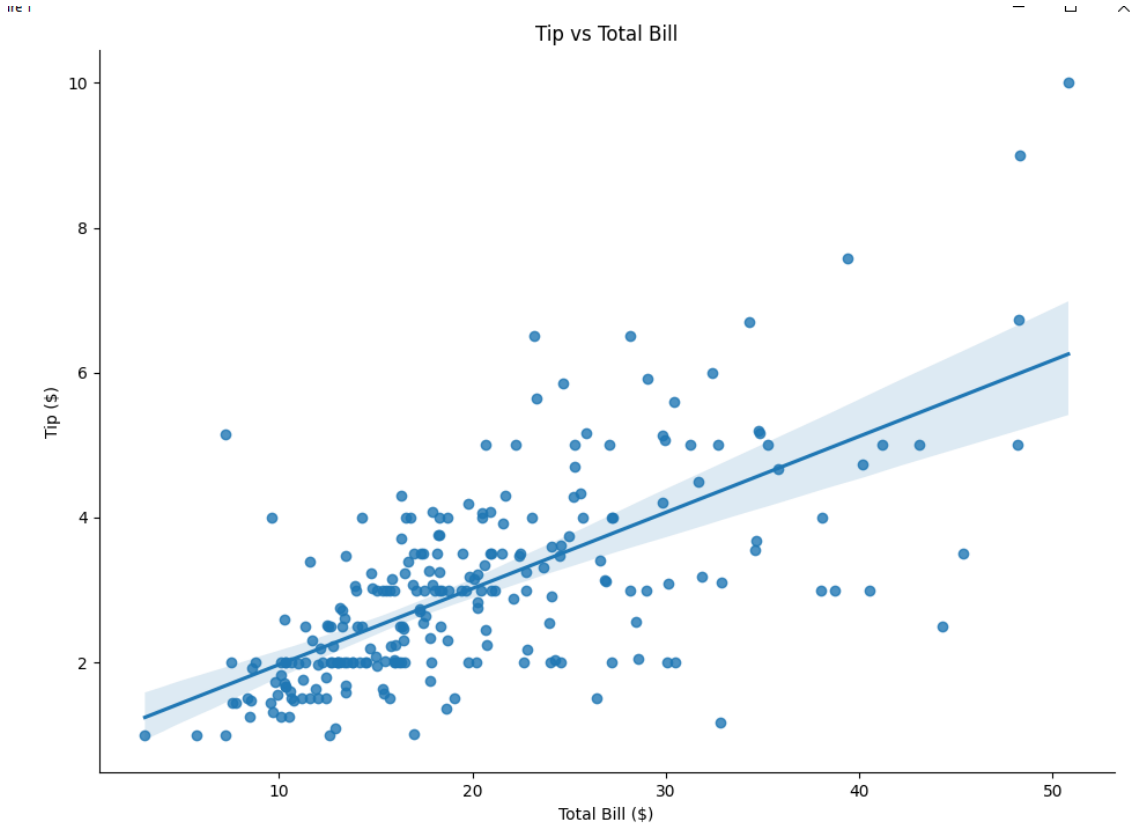
```
# Create DataFrames
bar_df = pd.DataFrame({'Categories': categories, 'Values': bar_data})
scatter_df = pd.DataFrame({'X': x_scatter, 'Y': y_scatter, 'Colors': scatte

# Create a figure with two subplots
fig, (ax1, ax2) = plt.subplots(1, 2, figsize=(12, 5))

# Bar plot using pandas
bar_df.plot(kind='bar', x='Categories', y='Values', color='skyblue', ax=ax1
ax1.set_title("Bar Plot", fontsize=16, fontweight='bold')
ax1.set_xlabel("Categories")
ax1.set_ylabel("Values")

# Scatter plot using pandas
scatter_df.plot(kind='scatter', x='X', y='Y', c='Colors', colormap='viridis
ax2.set_title("Scatter Plot", fontsize=16, fontweight='bold')
ax2.set_xlabel("X-axis")
ax2.set_ylabel("Y-axis")
```

# Seaborn

- Visualisatie library gebaseerd op Matplotlib
- Vaak gemakkelijker dan werken met Matplotlib
  - Een aantal speciale aanpassingen vereisen nog steeds matplotlib

# Seaborn-example

```python
import seaborn as sns
import matplotlib.pyplot as plt

# Load example dataset
tips = sns.load_dataset("tips")

# Create a scatter plot with regression line
sns.lmplot(x="total_bill", y="tip", data=tips)

# Customize the plot
plt.title("Tip vs Total Bill")
plt.xlabel("Total Bill ($)")
plt.ylabel("Tip ($)")

# Show the plot
plt.show()
```



# Zelfstudie

# Data visualization tutorial

- Ga naar:
  - https://www.kaggle.com/learn/data-visualization
  - Volg de tutorial volledig

  - De informatie in de tutorials is te kennen leerstof en helpt bij het maken van de oefeningen

# EDA oefening

- Werk verder aan de oefening
- Deze oefening wordt geëvalueerd