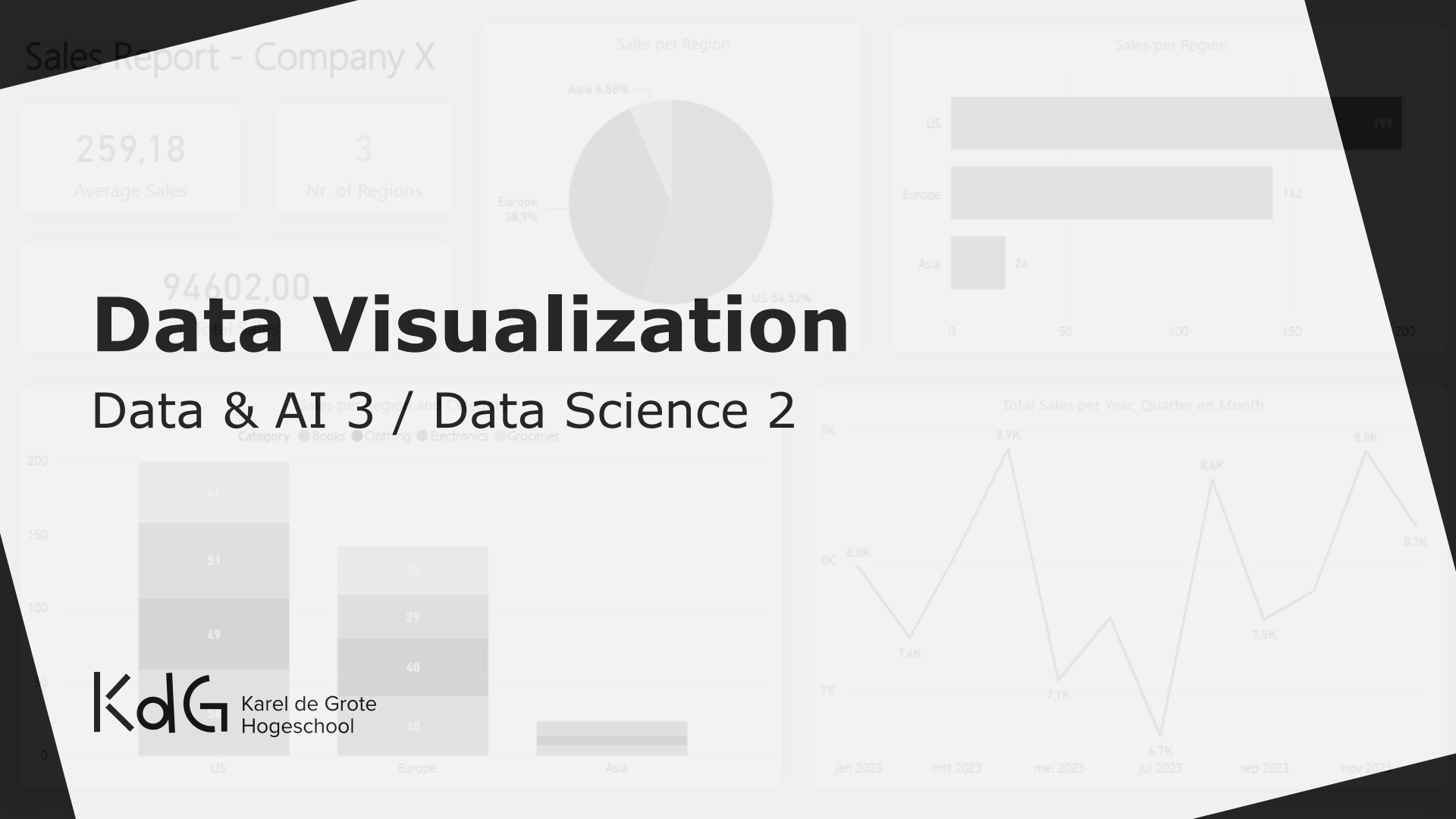
Data Visualization

Data & AI 3 / Data Science 2

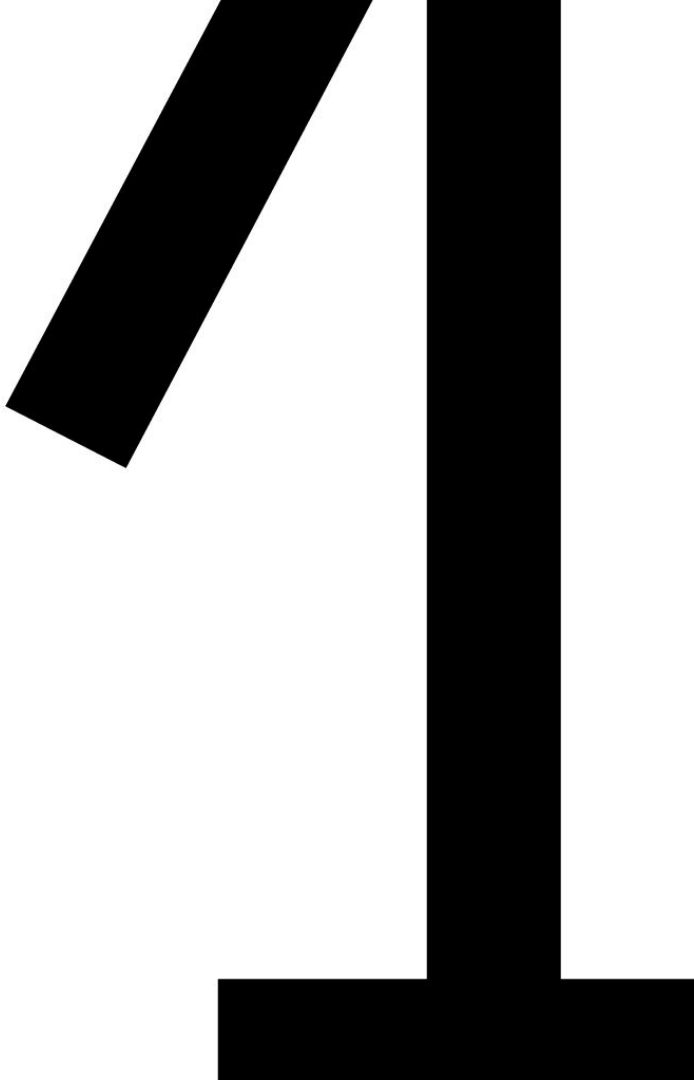


KdG Karel de Grote Hogeschool

Agenda

1. Introduction to Data visualization
2. Introduction to Matplotlib
3. Introduction to Seaborn
4. Types of Data Visualizations
5. Which Data Visualization to Choose

Introduction to data visualization



Intro to Data Visualization

- Purpose of Data Visualization:
  - Transforming data into a visual context.
  - Making data accessible, understandable, and actionable.

- Importance in Data Analysis:
  - Identifies trends, patterns, and outliers.
  - Simplifies complex data sets.

2

Introduction to

Intro to Matplotlib

- What is Matplotlib?
  - A powerful plotting library in Python.
  - Provides comprehensive control over plot elements.

- Key Features:
  - Highly customizable plots.
  - Large variety of plot types.

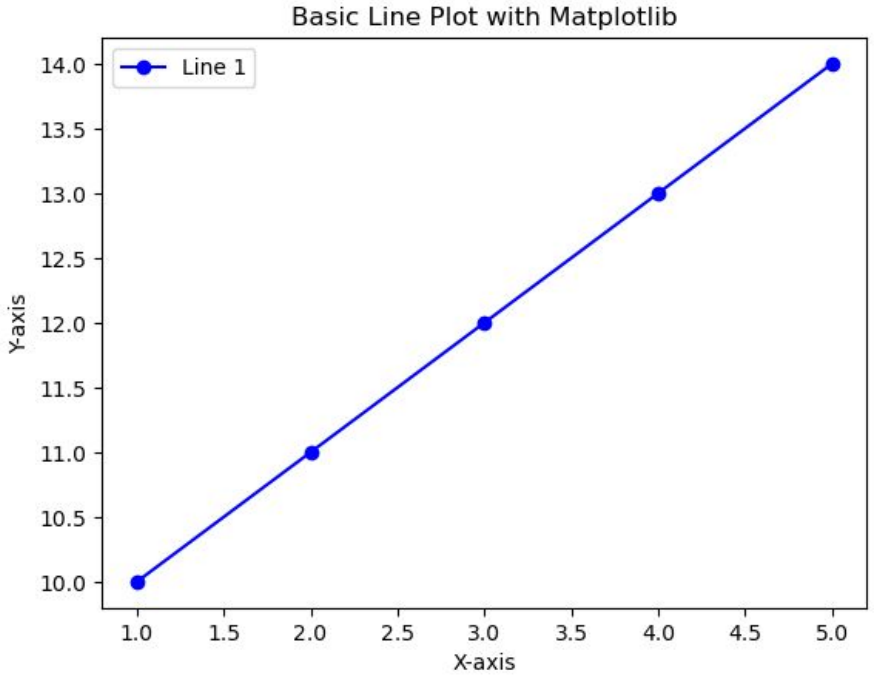
Intro to Matplotlib

Setting up Matplotlib

- Installation: pip install matplotlib
- Import: import matplotlib.pyplot as plt

Intro to Matplotlib

```python
# Importing the necessary libraries
import matplotlib.pyplot as plt

# Sample data
x = [1, 2, 3, 4, 5]
y = [10, 11, 12, 13, 14]

# Creating the line plot
plt.plot(x, y, label='Line 1', color='blue', marker='o')

# Adding title and labels
plt.title('Basic Line Plot with Matplotlib')
plt.xlabel('X-axis')
plt.ylabel('Y-axis')

# Displaying a legend
plt.legend()

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



Intro to Matplotlib

Notebook

See 04.00-Introduction-To-Matplotlib.ipynb

See 04.01-Simple-Line-Plots.ipynb

Exercise Time!

See 04.00_EX.ipynb

See 04.00_EX.ipynb

Introduction to

seaborn

Intro to Seaborn

- What is Seaborn?
  - Built on top of Matplotlib, focusing on statistical data visualization.
  - Provides an aesthetically pleasing interface and advanced plots.

- Key Features:
  - Themes for styling plots.
  - Simplified syntax for complex visualizations.

Matplotlib vs. Seaborn

- Matplotlib:
  - More control,
  - detailed customization.
- Seaborn:
  - Better default aesthetics,
  - easier to use for statistical plots.

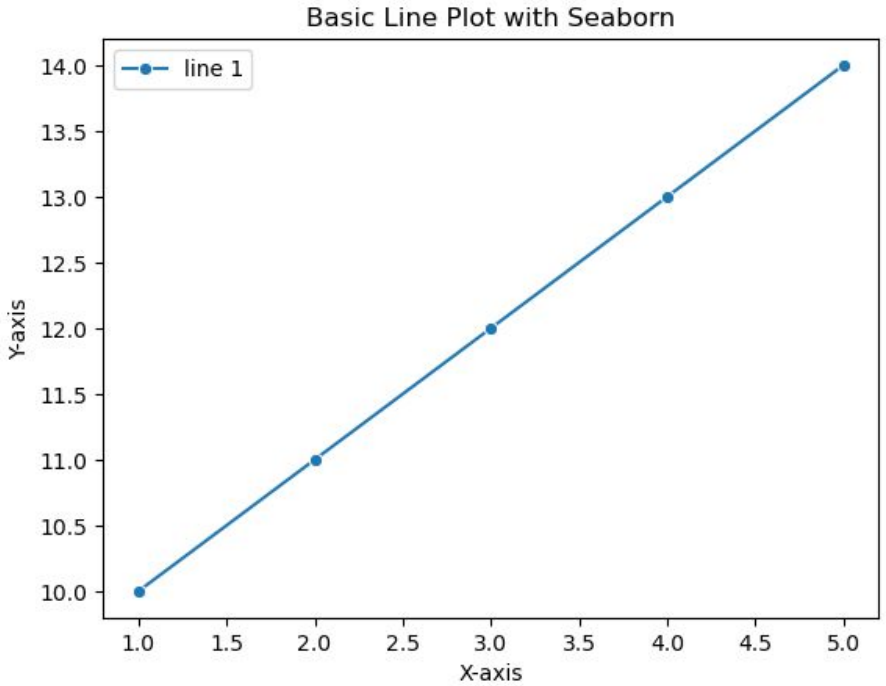
Intro to Seaborn

Setting up Seaborn:

- Installation: pip install seaborn
- Import: import seaborn as sns

Intro to Seaborn

```python
# Importing the necessary libraries
import seaborn as sns
import matplotlib.pyplot as plt

# Sample data
x = [1, 2, 3, 4, 5]
y = [10, 11, 12, 13, 14]

# Creating the line plot using Seaborn
sns.lineplot(x=x, y=y, marker='o', label='line 1')

# Adding title and labels
plt.title('Basic Line Plot with Seaborn')
plt.xlabel('X-axis')
plt.ylabel('Y-axis')

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



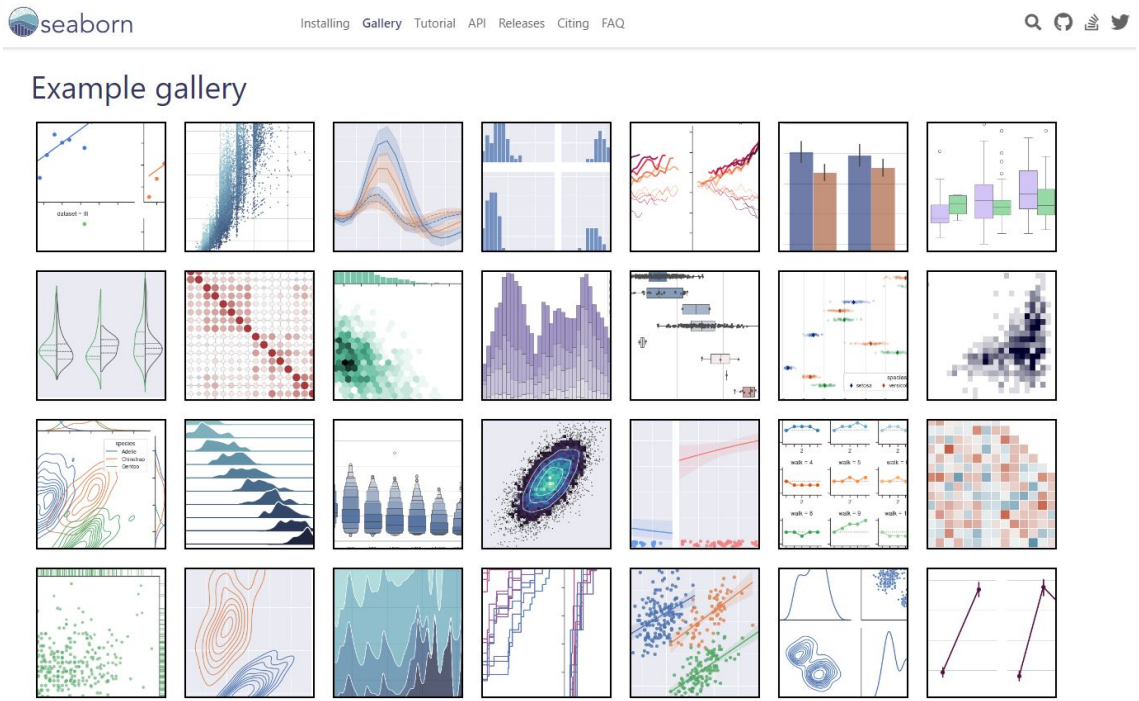
Types of Data Visualizations

Types of Data visualizations

- Bar chart/Stacked Barchart
- Histogram
- Line chart
- Pie chart
- Scatter plot
- Boxplot

Types of Data visualizations

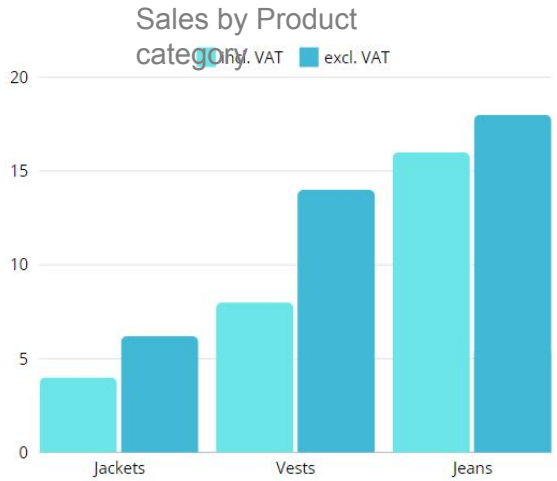
- Bar chart
- Histogram
- Line chart
- Pie chart
- Scatter plot
- Boxplot



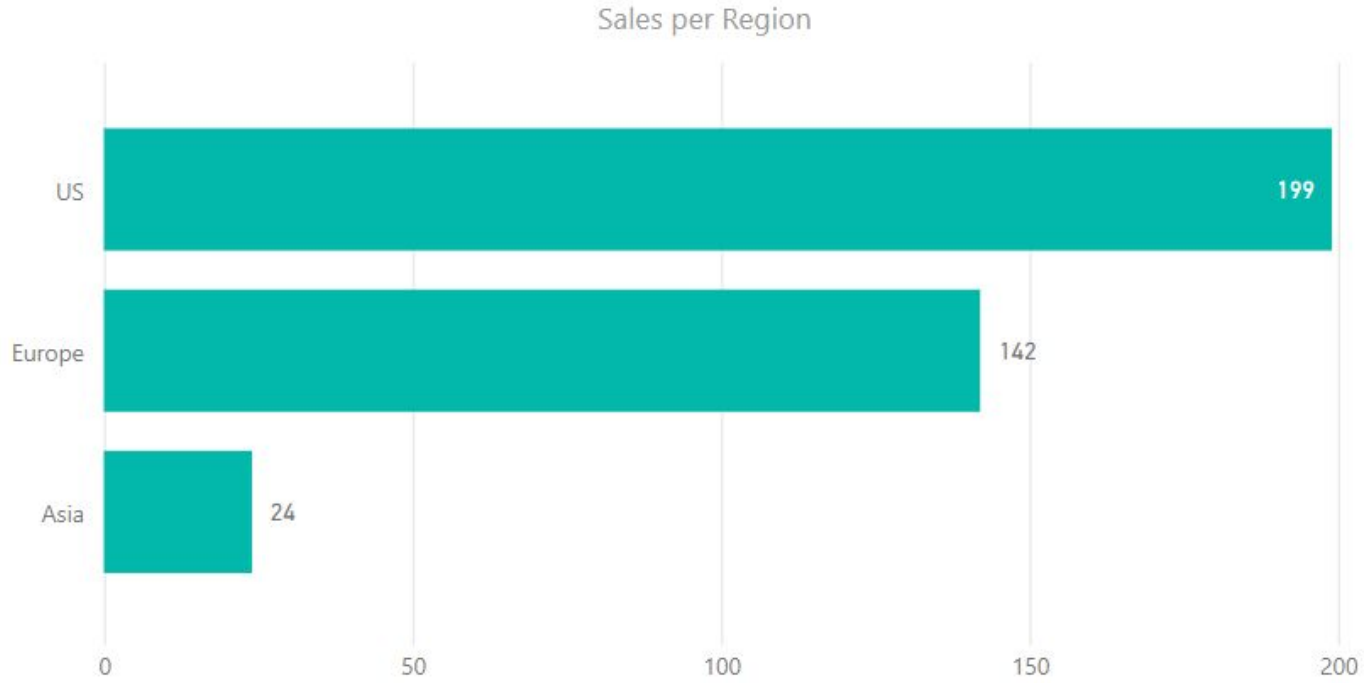
https://seaborn.pydata.org/examples/index.html

Graphs – Bar Chart

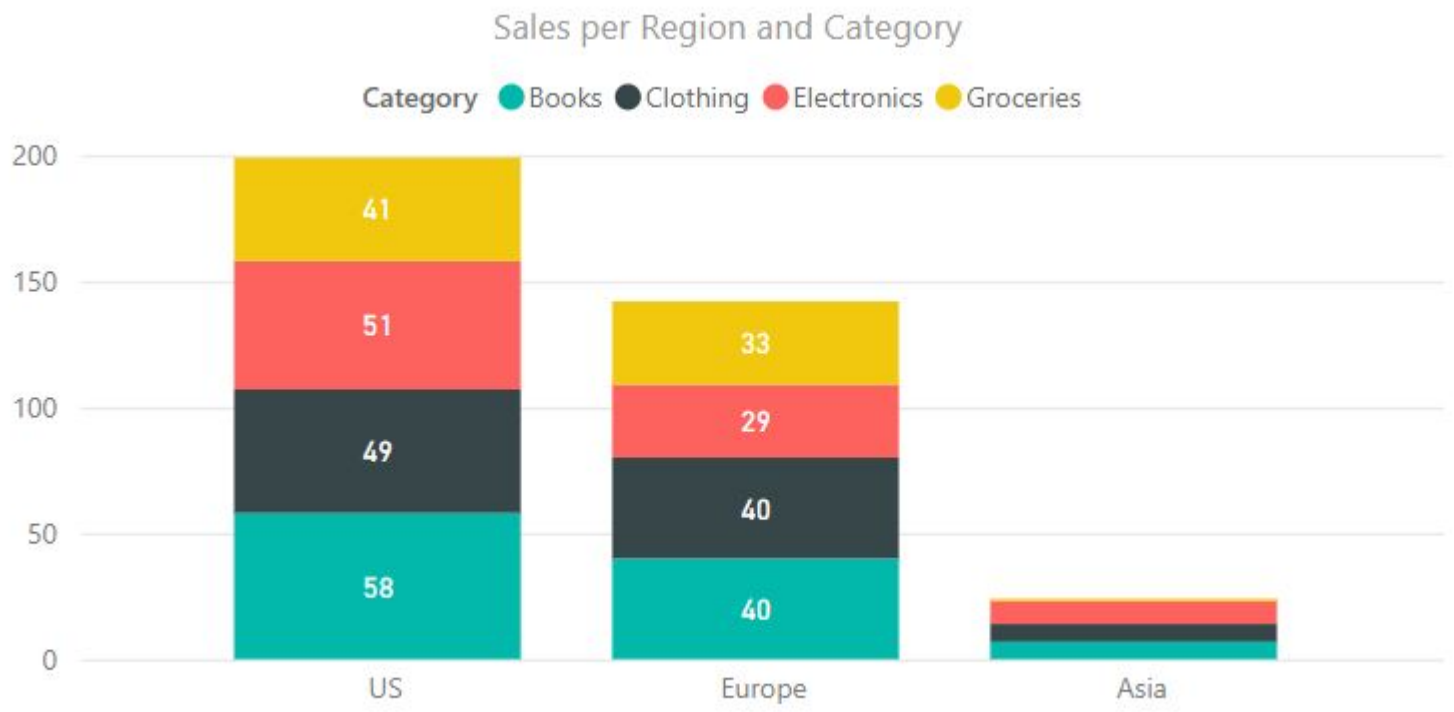
- What is it?
  - Shows data as bars that represent the height of categories
- When to use it?
  - Comparing discrete categories (for example sales per product)
  - Suited for nominal and ordinal data
- Example:
  - Number of sold products per product category.



Graphs – Bar Chart

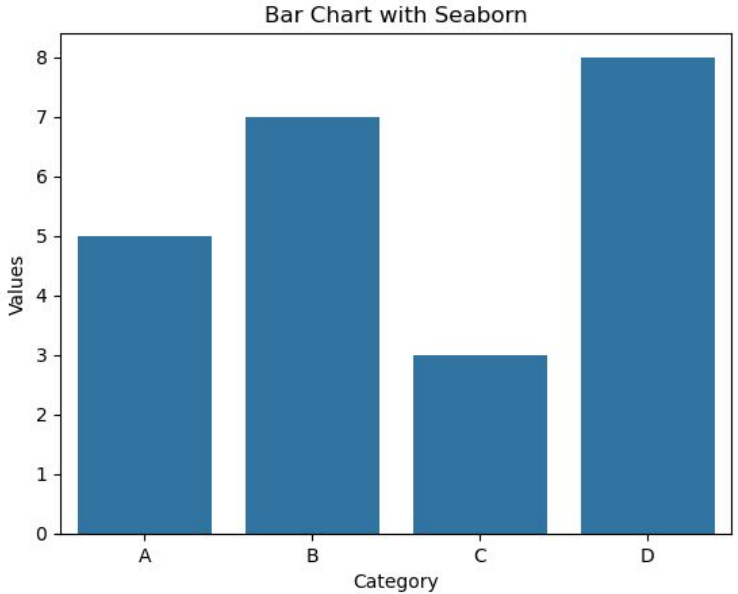


Graphs – Stacked Bar Chart



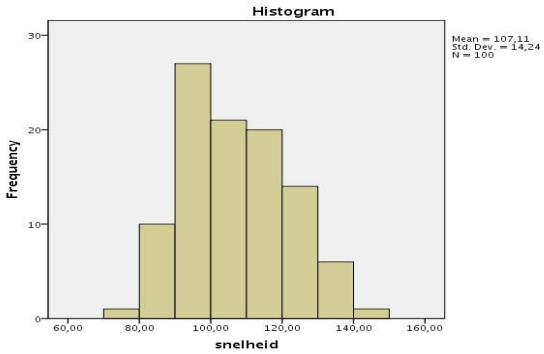
Bar Chart - Seaborn

```python
# Importing necessary libraries
import seaborn as sns
import matplotlib.pyplot as plt

# Sample data
categories = ['A', 'B', 'C', 'D']
values = [5, 7, 3, 8]

# Creating a bar plot using Seaborn
sns.barplot(x=categories, y=values)

# Adding a title and labels
plt.title('Bar Chart with Seaborn')
plt.xlabel('Category')
plt.ylabel('Values')

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



https://seaborn.pydata.org/examples/part_whole_bars.html

Graphs – Histogram



- What is it?
  - A graph that displays the frequencies of values in certain intervals
- When to use it?:
  - Visualization of the distribution of continuous variables
  - Suitable for interval- or ratio-data
- Visualization = Bar chart
  - But with the bars connected to each other!
- Example:
  - Distribution of ages in a certain population

Graphs – Histogram

Notebook

See 04.05-Histograms-and-Bi nnings.ipynb

Exercises

04.01_EX.ipynb

Seaborn Histogram:

https://seaborn.pydata.org/examples/histogram_stacked.html
https://seaborn.pydata.org/examples/faceted_histogram.html

Graphs – Line Chart

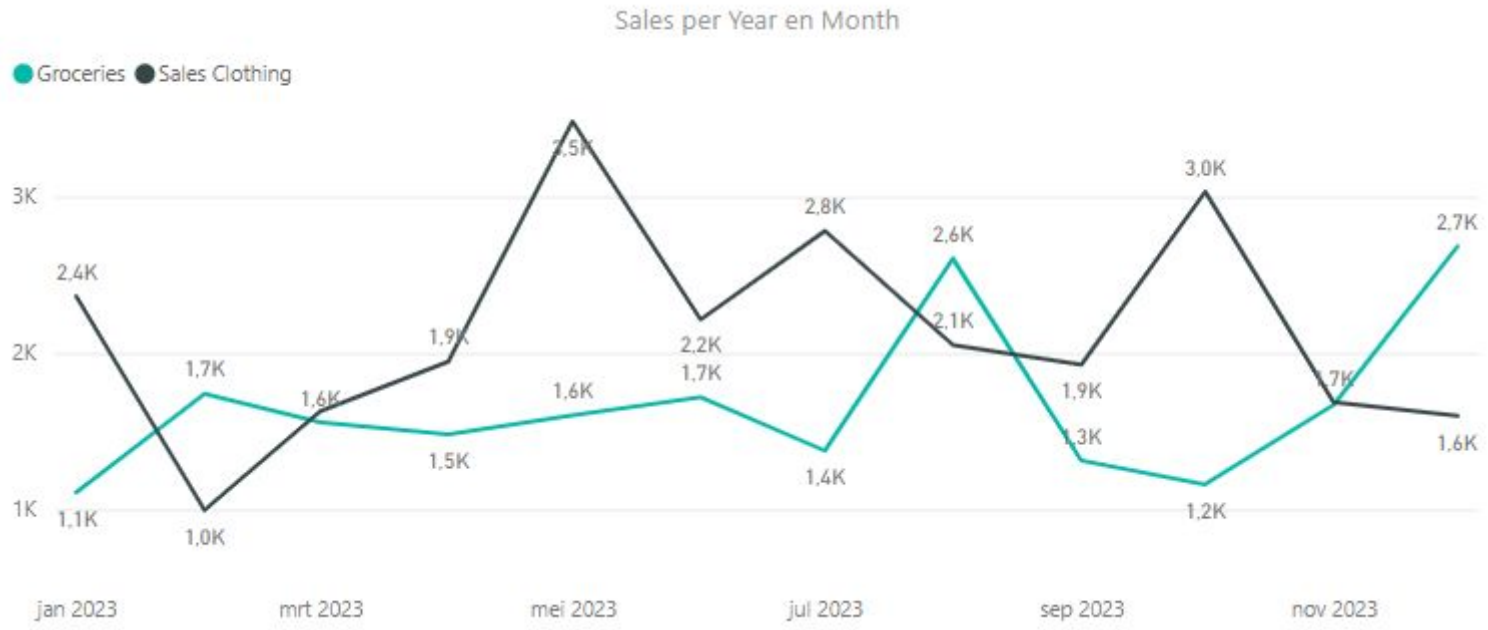
- What is it?
  - A line that shows the relation between two variables, mostly between time and another variable
- When to use it?
  - Showing trends over time (timeseries)
  - Continuously monitoring data (for example: Temperature changes)
- Example:
  - Sales over the months of a year

Graphs – Line Chart

Correct use:

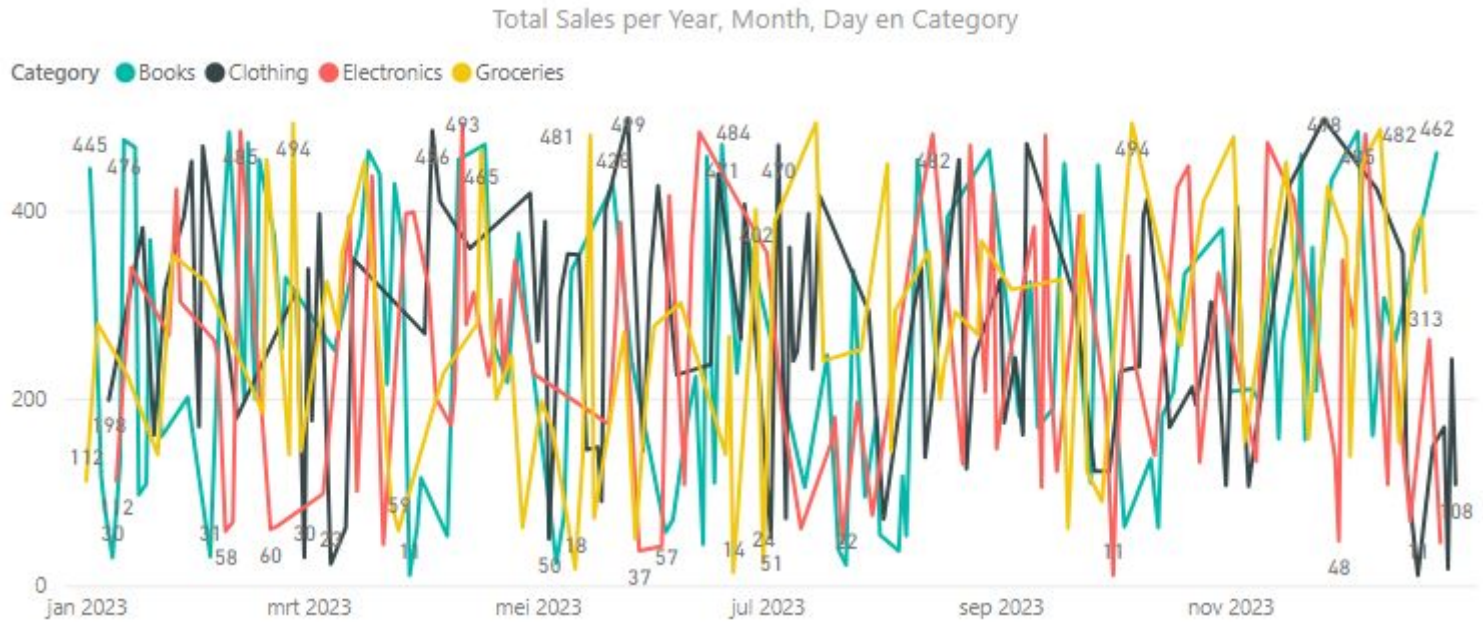
Graphs – Line Chart

Correct use:



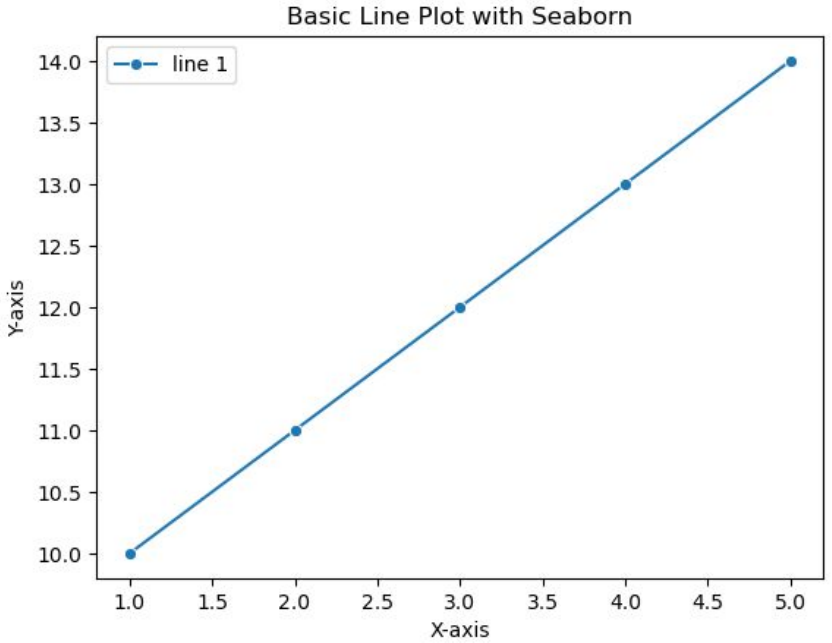
Graphs – Line Chart

Incorrect use:



Line Chart with Seaborn

```python
# Importing the necessary libraries
import seaborn as sns
import matplotlib.pyplot as plt

# Sample data
x = [1, 2, 3, 4, 5]
y = [10, 11, 12, 13, 14]

# Creating the line plot using Seaborn
sns.lineplot(x=x, y=y, marker='o', label='line 1')

# Adding title and labels
plt.title('Basic Line Plot with Seaborn')
plt.xlabel('X-axis')
plt.ylabel('Y-axis')

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



https://seaborn.pydata.org/examples/errorband_lineplots.html

Graphs - Pie Chart

- What is it?
  - Shows parts of a whole in circlesegments.
- When to use it?
  - Showing distributions as part of a whole (100%)
  - Suited for relatively few categories (ideally maximum 5)

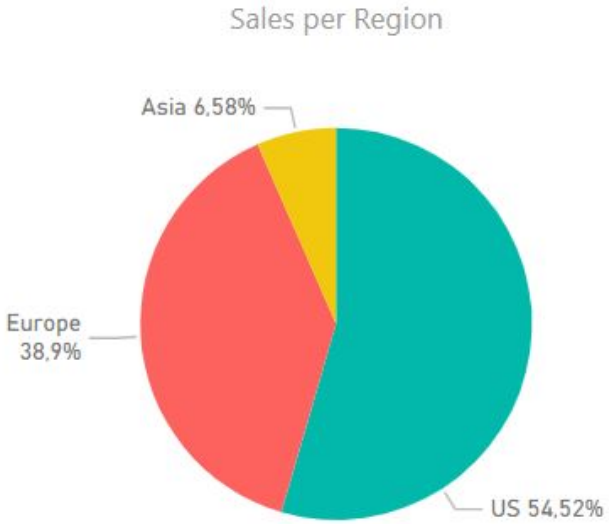
Example:

  - Marketshare of 3 different companies.

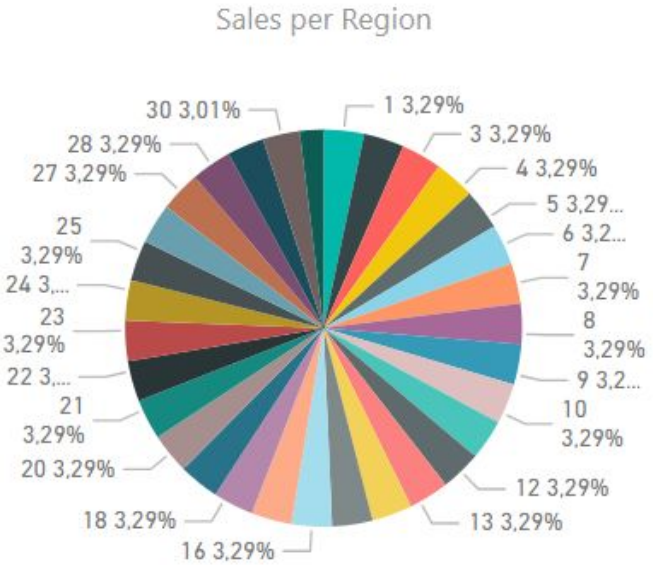
- **Remark**: Not suited for many categories or if the exact differences between categories are important!

Graphs - Pie Chart

Correct use:

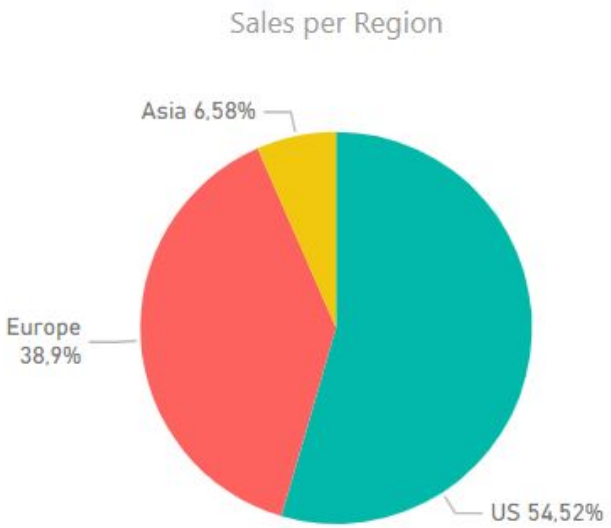


Incorrect use:

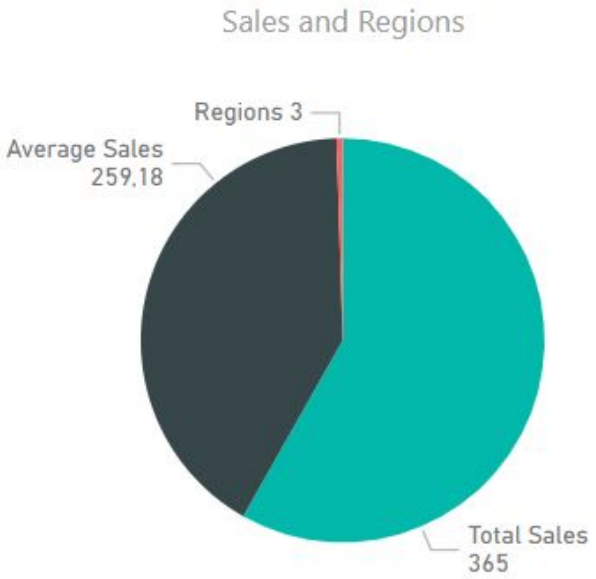


Graphs - Pie Chart

Correct use:



Incorrect use:



Graphs – Scatter Plot

- What is it?:
  - Show relationship between two variables using dots.
- When to us it?
  - Relationship or correlation between two variables
  - Shows patterns like linear or non-linear relationships
- Example:
  - Relationship between length and weight.

Graphs – Scatter Plot

Graphs – Scatter Plot

Notebook

See 04.02-Simple-Scatter-Plots.ipynb

Exercises

04.00_EX.ipynb

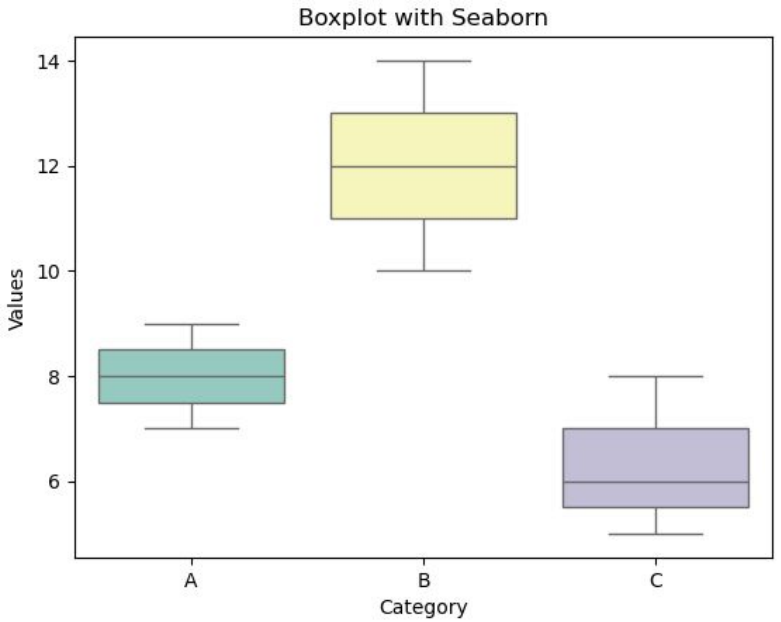
Seaborn Scatter plot:
https://seaborn.pydata.org/examples/different_scatter_variables.html

Graphs – Boxplot

- What is it?
  - Shows dispersion of a dataset using the min, Q1, median, Q3, max values.
- When to use it?
  - Visualisation of dispersion and outliers
  - To compare different datasets.
- Example:
  - Compare the dispersion of salaries between different sectors.

Graphs – Boxplot in Seaborn

```python
# Importing necessary libraries
import seaborn as sns
import matplotlib.pyplot as plt
import pandas as pd

# Sample dataset in the form of a pandas DataFrame
data = {
    'Category': ['A', 'A', 'A', 'B', 'B', 'B', 'C', 'C', 'C'],
    'Values': [7, 8, 9, 12, 10, 14, 6, 5, 8]
}

df = pd.DataFrame(data)

# Creating a boxplot using Seaborn
sns.boxplot(x='Category', y='Values', data=df, palette='Set3')

# Adding a title and labels
plt.title('Boxplot with Seaborn')
plt.xlabel('Category')
plt.ylabel('Values')

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



https://seaborn.pydata.org/examples/grouped_boxplot.html

Which Data Visualization to Choose and When

Graphs – frequent mistakes

- Too many categories in a pie or line chart.
- Wrongly chosen axes (manipulating the Y-axes)
- Wrong graph for the type of data

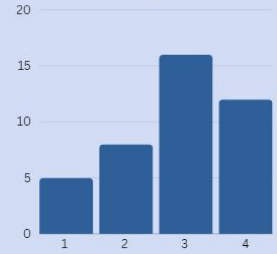
<u>Tips to make graphs better and more effective:</u>

- Only show the information that is necessary
- Don’t put too much information in 1 graph, it is better to spread the information over multiple graphs instead.

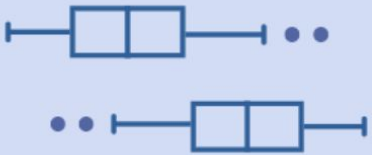
What do you want to show?

How the data is distributed

Frequency distribution:
Histogram

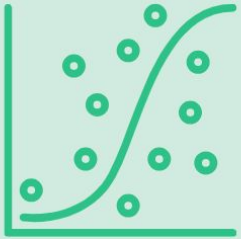


Statistical dispersion:
Boxplot



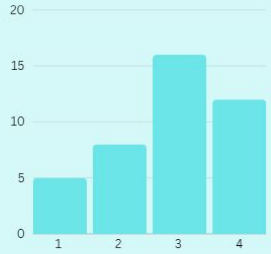
How values relate to each other

Scatterplot



How values compare to each other

Bar chart



Pie chart

Over time: Line chart

Graphs – When to use which one?

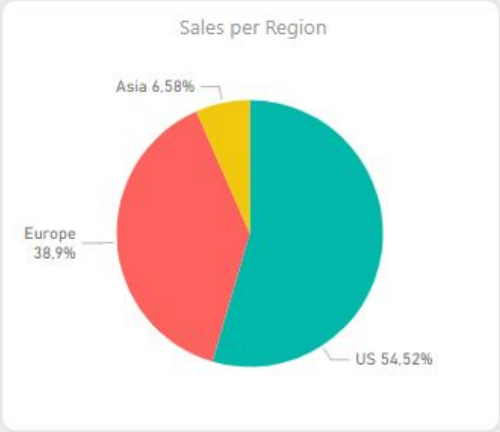
- Bar chart vs. Pie chart:
  - Use the bar chart when it is important to see exact differences between categories.
  - Use Piecharts only for up to 5 different categories
- Linechart vs. histogram:
  - use histogram for distributions, linecharts for trends over time (time series).
- Boxplot vs. scatter plot:
  - Use a boxplot for dispersion and a scatterplot for relationships between to variables.

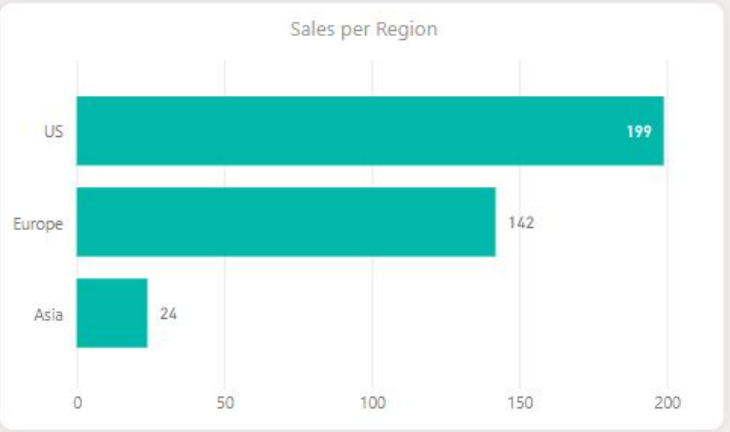
Sales Report - Company X

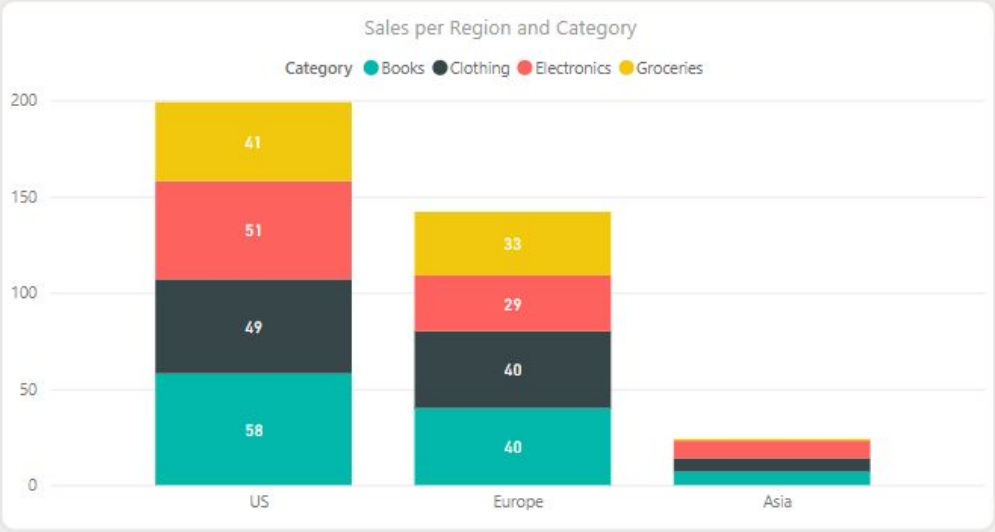
259,18
Average Sales

3
Nr. of Regions

94602,00
Total Sales







Graphs – Conclusion

- Know when to use which graph
- Graphs should make it easier to understand data
- Avoid making graphs with high complexity

Graphs – Conclusion

Notebook

See 04.06-Customizing-Legends.ipynb

See 04.07-Customizing-Colorbars.ipynb

See 04.14-Visualization-With-Seaborn.ipynb

Exercises

04.02_EX.ipynb