Data Visualization With Pandas

Overview

In this lesson, students will learn how to use the Pandas library to visualize data.

Duration

120 minutes

Learning Objectives

In this lesson, students will:

- Explain the characteristics of a great data visualization.
- Identify when to use a bar chart, pie chart, line chart, scatterplot, or histogram.
- Use Pandas to implement line charts, bar charts, scatterplots, and histograms.

Pre-Class Materials and Preparation

For remote classrooms: Virtual breakout rooms and Slack may be needed to facilitate the partner exercise and discussions. As you plan for your lesson:

- Consider how you'll create pairs for the partner exercise (randomly, or with pre-assigned partners).
- Determine how (if at all) exercise timing may need to be adjusted.
- For helpful tips, keep an eye out for the **For remote classrooms** tag in the speaker notes.
- Prepare screenshots and answers to exercises in advance so that they can be easily shared in Slack during your lecture.

Suggested Agenda

Time	Activity
0:00–0:20	**Welcome + Data Visualization Best Practices**
0:20–1:00	**Using Pandas to Visualize Data**
1:00–1:10	**Break**
1:10–1:50	**Using Pandas to Visualize Data (Cont.)**
1:50–2:00	**Wrapping Up, Q&A, and Exit Ticket Completion**

Jupyter Notebook

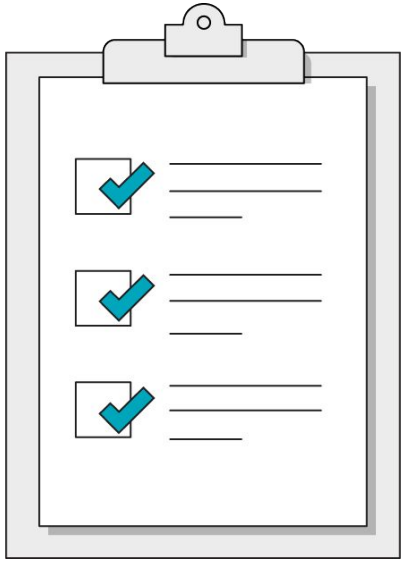
The exercises referred to in this lesson can be found in the Python Workbooks + Data folder.

Data Visualization With Pandas

GENERAL ASSEMBLY

Our Learning Goals

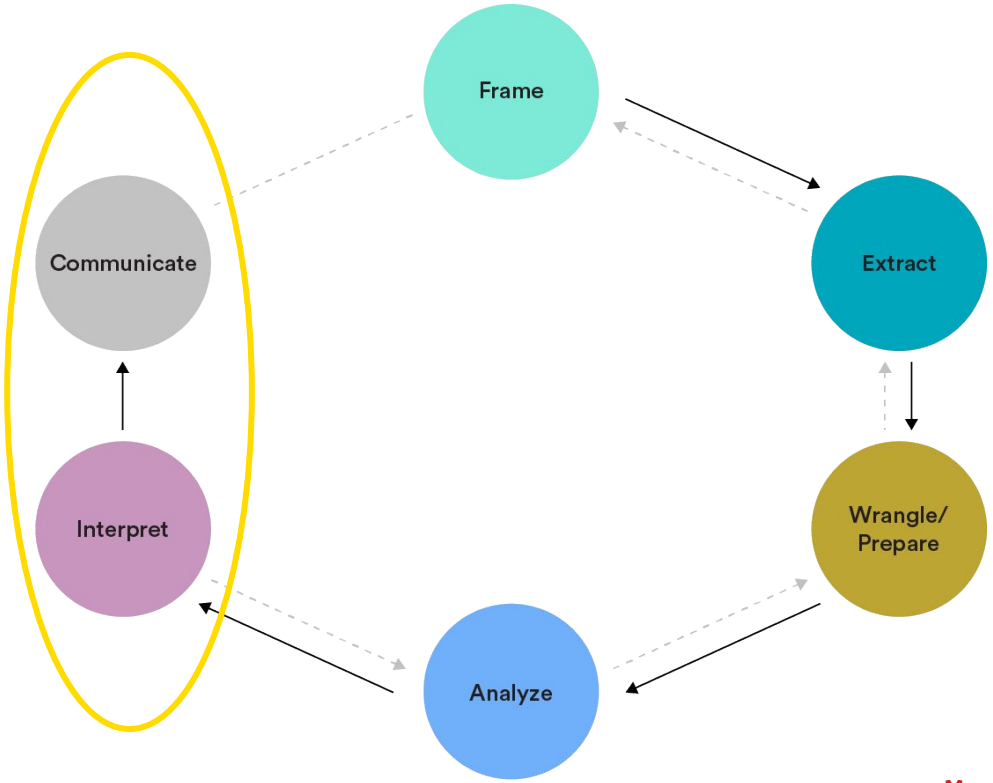
- Explain the characteristics of a great data visualization.
- Identify when to use a bar chart, pie chart, line chart, scatterplot, or histogram.
- Use Pandas to implement line charts, bar charts, scatterplots, and histograms.



The Data Analytics Workflow

Interpret: Leverage your analysis to make decisions and recommendations.

Communicate: Present data-driven findings and insights in a compelling manner.



Data Visualization With Pandas

Data Visualization Best Practices

“

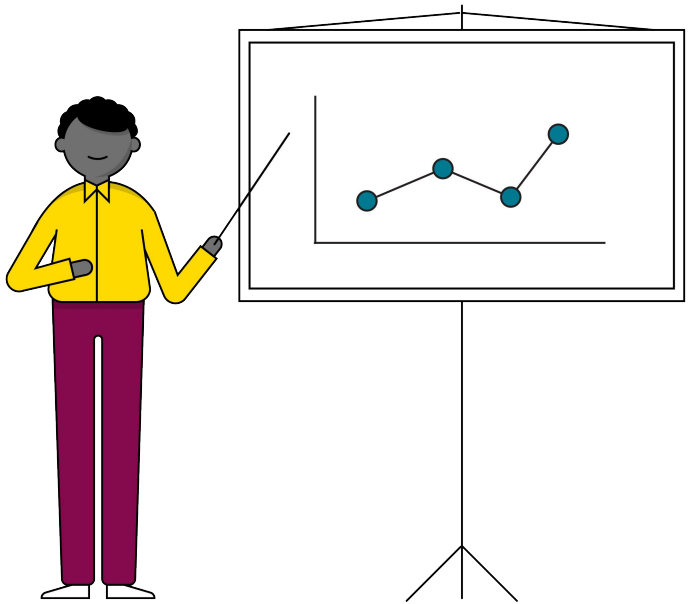
Use a picture. It’s worth a thousand words.

— Tess Flanders, newspaper editor, 1911

What Is Data Visualization?

Data visualizations are a quick way to convey insights from data sets. Specifically, we can use visuals to:

- Identify outliers during exploratory data analysis.
- Describe patterns or trends in the data.
- Summarize the results of an analysis and communicate key insights.



Criteria for Good Visualizations

1. **Simplified**
   Distill your insight to its essential components.

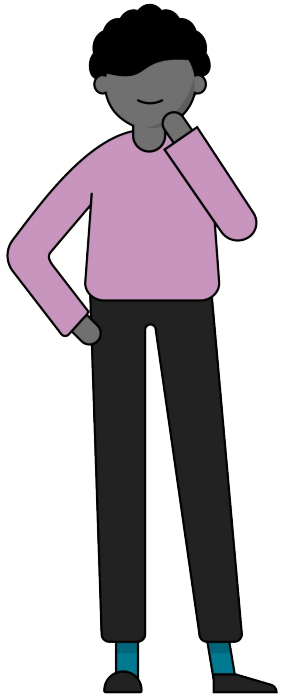
2. **Easy to Interpret**
   Use logical ordering and understandable metrics.

3. **Clearly Labeled**
   Use clearly distinguishable colors and avoid cluttering text.

Choosing the Right Chart Type

With so many chart types, it can be difficult to know how best to display your data.

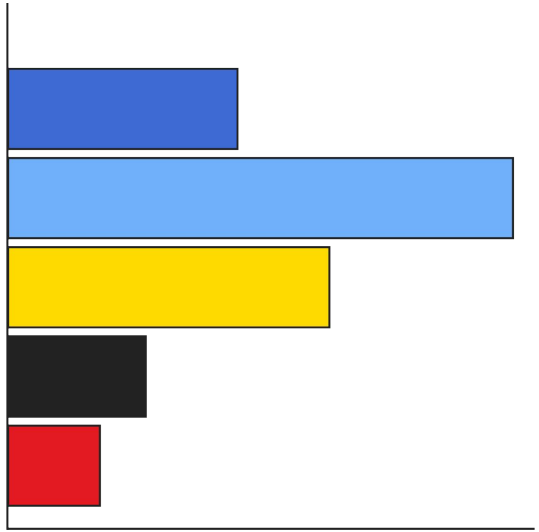
When creating a visualization, first think about the **variables** you are displaying, the **volume** of data you're showcasing, and the **central point you are hoping to communicate** through your visualization.



Bar Charts

Bar charts are one of the most common ways of visualizing data. Why? Because they make it easy to **compare information among categories**, revealing highs and lows quickly and efficiently.

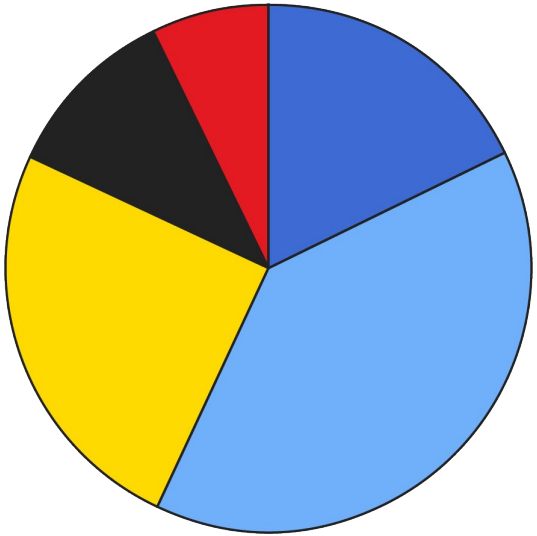
Bar charts are most effective when you have categorical data and want to explore counts of different categories.



Pie Charts

Pie charts are the most commonly misused chart type. If there are more than 2–3 categories involved, they can often mislead readers. Pie charts should only be used to show **relative proportions or percentages** of information.

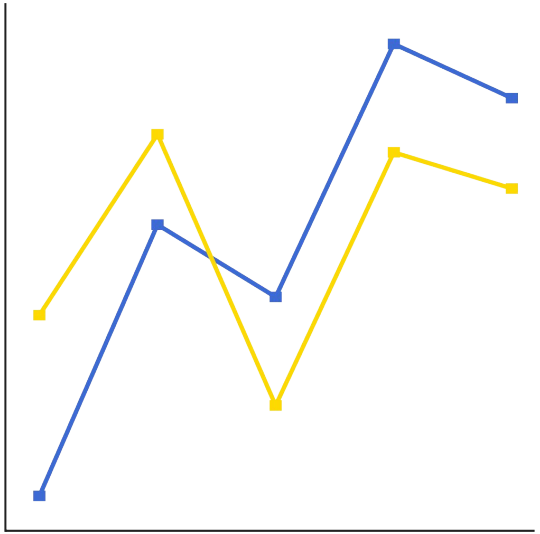
If you want to compare data, leave it to bars or stacked bars.



Line Charts

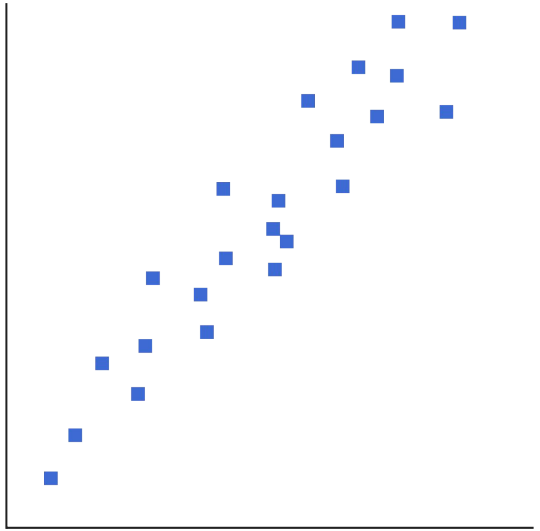
Line charts (or line graphs) are an excellent way to show **change over time**. They work best when you have one time variable and one set of data points to show.

While bar charts can also display time, they don’t show it in a continuous way.



Scatterplots

Scatterplots are great for giving viewers a sense of trends, concentrations, and outliers. They're also useful for illustrating **correlations** between two variables, depending on whether the shape is clearly concentrated or scattered far apart. This will provide a clear idea of what you may want to investigate further.



Knowledge Check!

Which type of data visualization should we use for the scenario below?

Scenario: Change in average income since 1960 for American adults.

Answer: Line charts are ideal for expressing change over time.

Knowledge Check!

Which type of data visualization should we use for the scenario below?

Scenario: Amount of sales per state.

Answer: Bar charts are best for comparing numbers.

Knowledge Check!

Which type of data visualization should we use for the scenario below?

Scenario: Determine if there's a correlation between book length and sales.

Answer: Scatterplots can compare the relationship between two variables.

Data Visualization With Pandas

Using Pandas to Visualize Data

Pandas ❤️ Matplotlib

Pandas DataFrame objects use another library, known as Matplotlib, behind the scenes.

This means you can use Matplotlib functions in combination with Pandas methods to alter plots after drawing them.

For example, you can use Matplotlib's **`xlabel()`** and **`title()`** functions to label the plot's x axis and title, respectively, after it is drawn.

The Plot Thickens

As we explore different types of plots, notice:

1. **Different types of plots are drawn very similarly** — they even tend to share parameter names.
2. In Pandas, **calling `plot()` on a DataFrame is different than calling it on a Series**. Although the methods are both named "plot," they may take different parameters.

Our First Chart

Once you've loaded data into a Pandas DataFrame, creating a chart is as simple as using the **`.plot()`** method.

```
import pandas as pd

import matplotlib.pyplot as plt

data_frame = pd.read_csv(file_address)

data_frame['column_name'].plot()
```

Plot Parameters

You may want to alter certain aspects of the chart, such as:

- The **kind** of plot you want (line, bar, scatter, etc.).
- The **style** of the lines, including color and line consistency.
- The size of the chart, or **`figsize`**.
- ... plus many other settings.

Customizations can be made using keyword parameters:

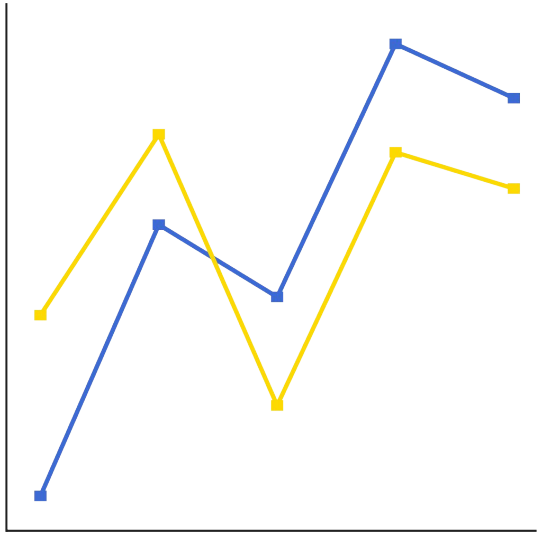
```
data_frame['column_name'].plot(style={'col1': 'r'}, figsize=(16,9))
```

Guided Walk-Through:

9.1 Line Charts in Pandas

10 minutes

Let's practice creating line charts in Section 9.1 of the workbook.



Discussion:

Counting Games per Country

If we alter our chart from using the “year” column to using the “country” column, all of a sudden the line chart stops making sense.

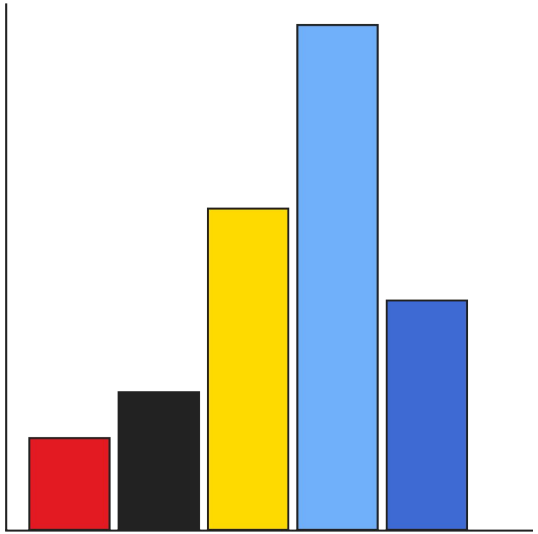
What chart should be used instead to compare the amount of games per country?

Partner Exercise:

9.2 Bar Charts

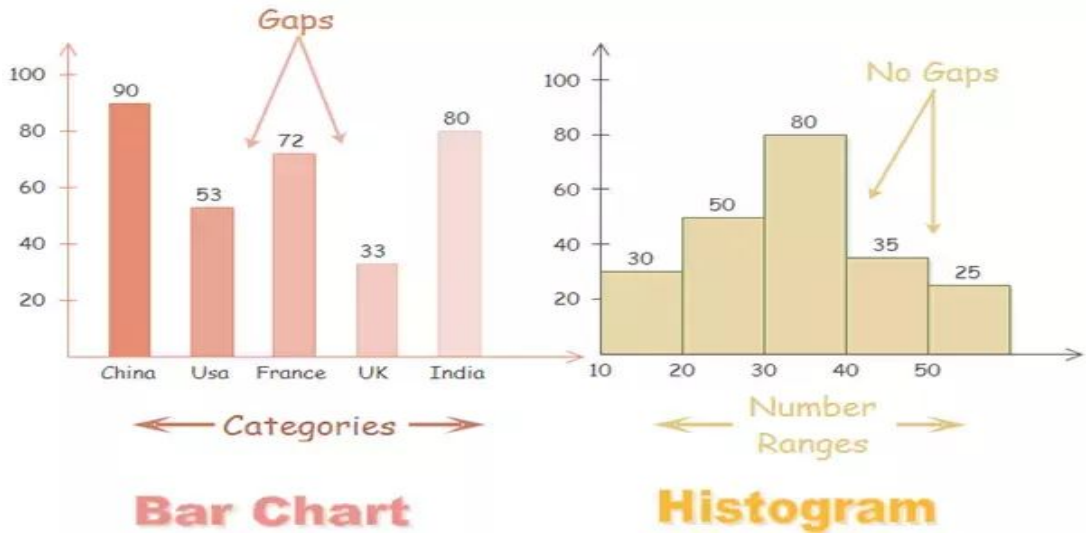
15 minutes

Let's use the same data set to start creating bar charts in Section 9.2.



Bar Charts vs. Histograms

Another common chart style is a **histogram**, which plots the distribution of values according to numerically defined groups rather than distinct categories.

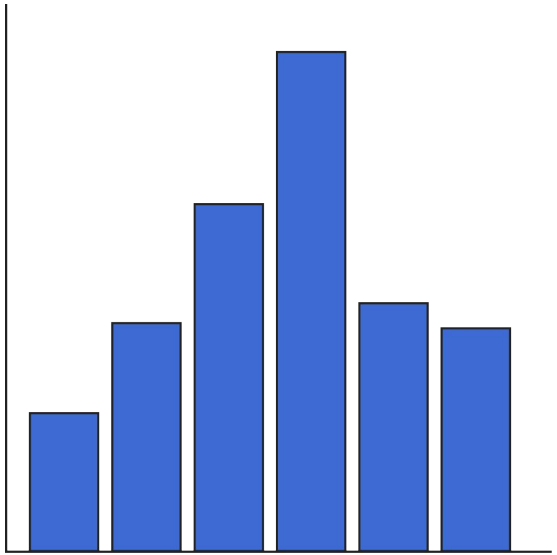


Source: Edrawsoft

Guided Walk-Through:
9.3 Histograms

10 minutes

Let’s look at some of the challenges of histograms in Section 9.3.



Scatterplots

Scatterplots intend to demonstrate the correlation, or lack thereof, between different variables. Therefore, we have to specify which columns to compare:

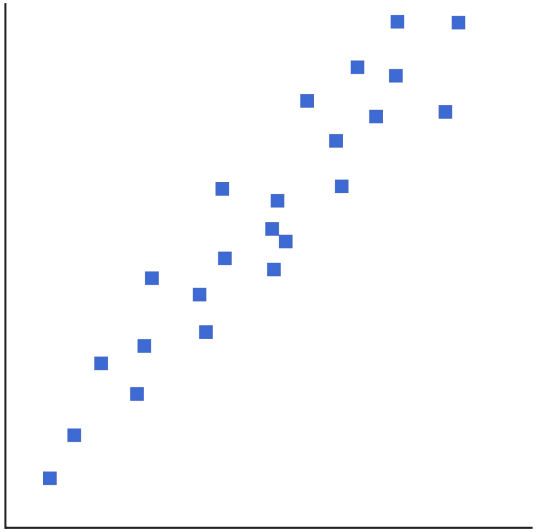
```
data_frame.plot(kind='scatter', x='column_a', y='column_b')
```

Scatterplots are most useful when values are **continuous**, rather than discrete with large gaps.

Guided Walk-Through:
9.4 Scatterplots

10 minutes

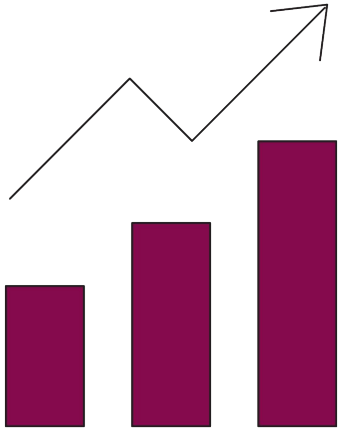
Let's practice using scatterplots to investigate correlations in Section 9.4.



Group Exercise:
9.5 Visual Storytelling With Data

25 minutes

Working with the Superstore data set, use **exploratory data analysis methods** and **at least one data visualization** to communicate trends, outliers, and a hypothesis surrounding the data.



Data Visualization With Pandas

Wrapping Up

Recap

In today's class, we...

- Explained the characteristics of a great data visualization.
- Identified when to use a bar chart, pie chart, line chart, scatterplot, or histogram.
- Used Pandas to implement line charts, bar charts, scatterplots, and histograms.

Looking Ahead

On your own:

- Work through the Python progress assessment on myGA (due at the end of the unit).
- Share your capstone project ideas with your instructor for review.
- Join someone else's project or invite others to join yours!

Next Class:

Cleaning and Combining Data With Pandas

Don’t Forget: Exit Tickets!


GA