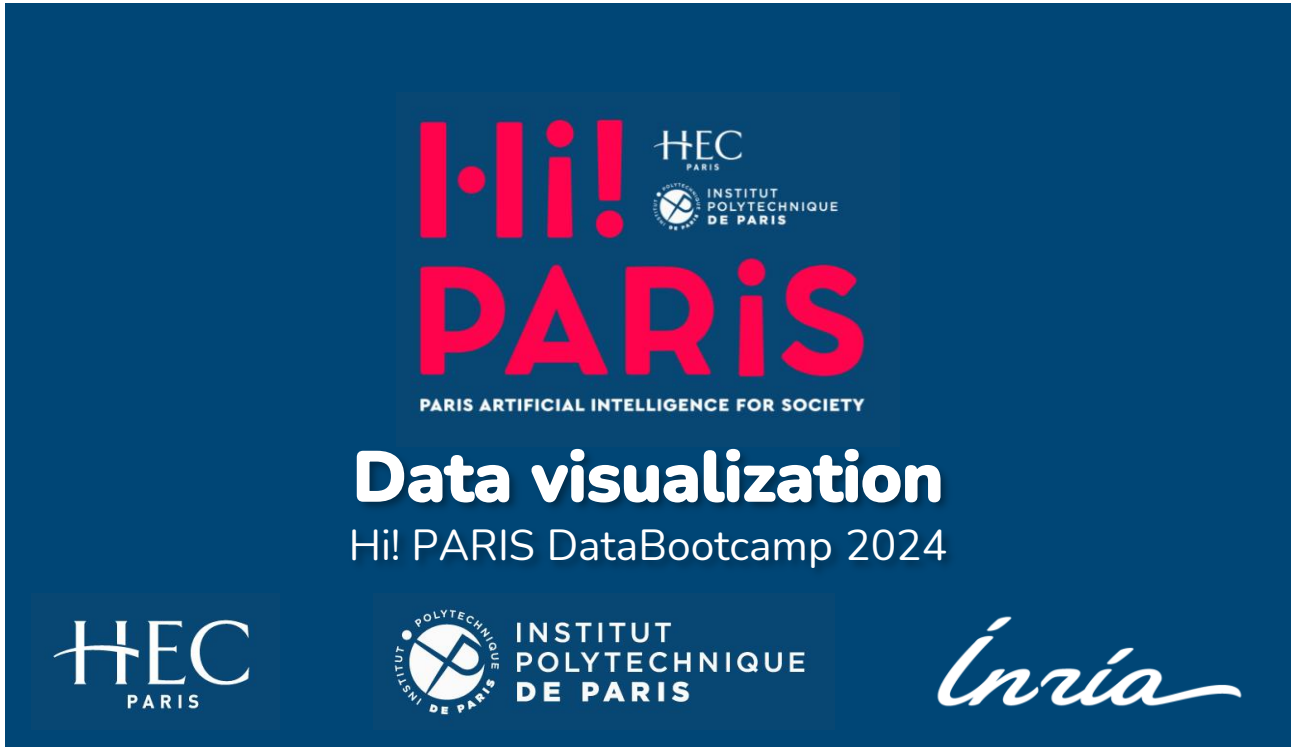


***Author**: Laurène DAVID, Machine Learning Research Engineer @ Hi! PARIS*

Agenda

I. What is Data Visualization ?

II. Main types of visualization plots

III. Exploratory Data Analysis (EDA)

What is Data visualization ?

What is Data visualization ?

Data visualization is the representation of data through the use of graphics, such as **charts/plots, maps,** and **interactive dashboards**

What is Data visualization ?

Data visualization is the representation of data through the use of graphics, such as **charts/plots, maps,** and **interactive dashboards**

1. **Identify Patterns and Trends**
   Detect trends or anomalies that aren't apparent when looking at raw data

2. **Easily communicate insights**
   Summarize data into visual representations that are easier to comprehend

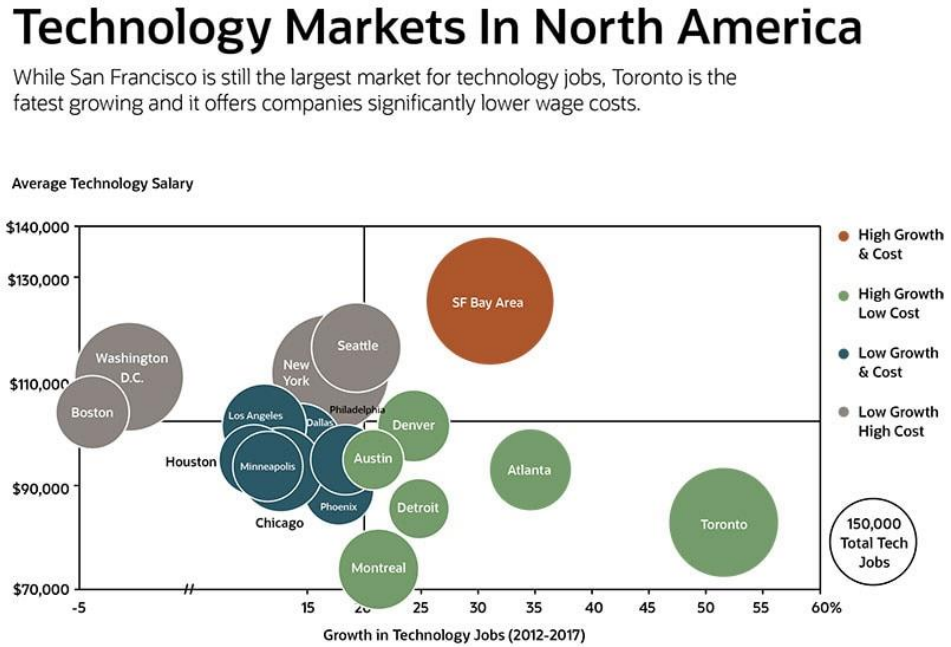
3. **Support Decision-Making**
   Absorb information quickly and make informed decisions based on the information provided.

What is Data visualization ?

Charts and plots

Summarize data in a **graphical form** and display it along **two or three axes.**

➡ *Data can be represented as* ***points, lines, curves , areas, bars****…*

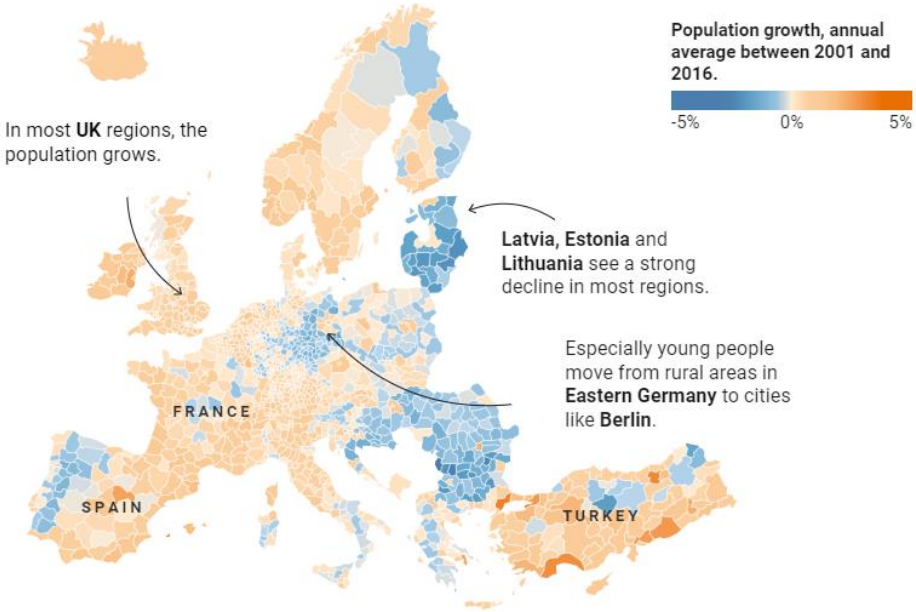


What is Data visualization ?

Maps and geospatial plots

Represent data on a **map**, with **geographical information** and **spatial relationships**.

➡ *Types of geospatial plots: choropleth, locator, symbol …*



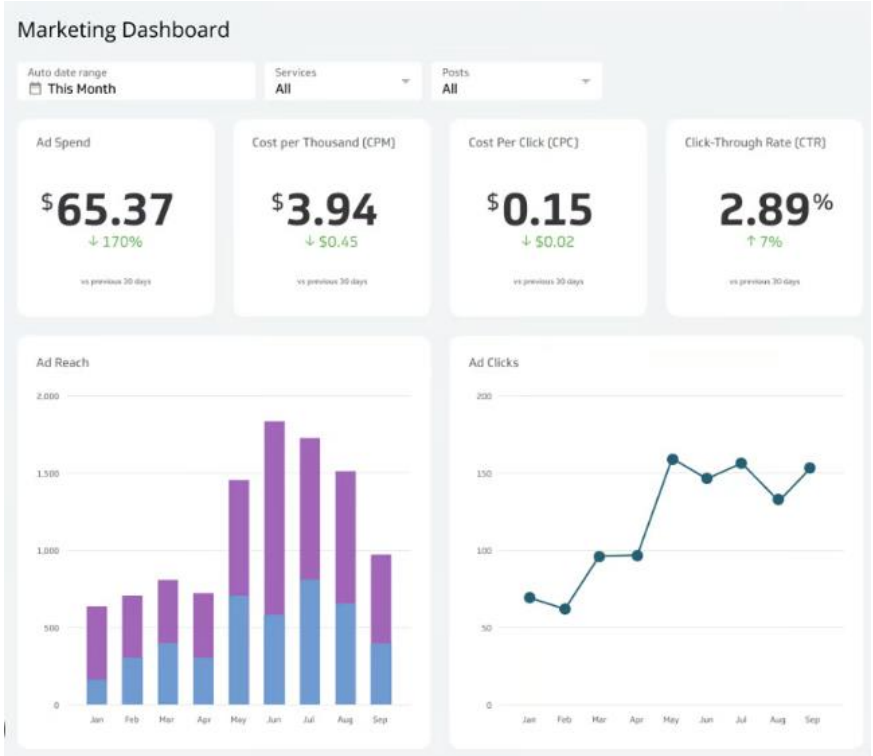
Population growth per region with a choropleth map

What is Data visualization ?

Interactive dashboards

Collection of charts that summarize **key indicators** and display data **interactively**

- *Get a real-time snapshot of KPIs and other important metrics.*
- *Users can personalize the information shown*



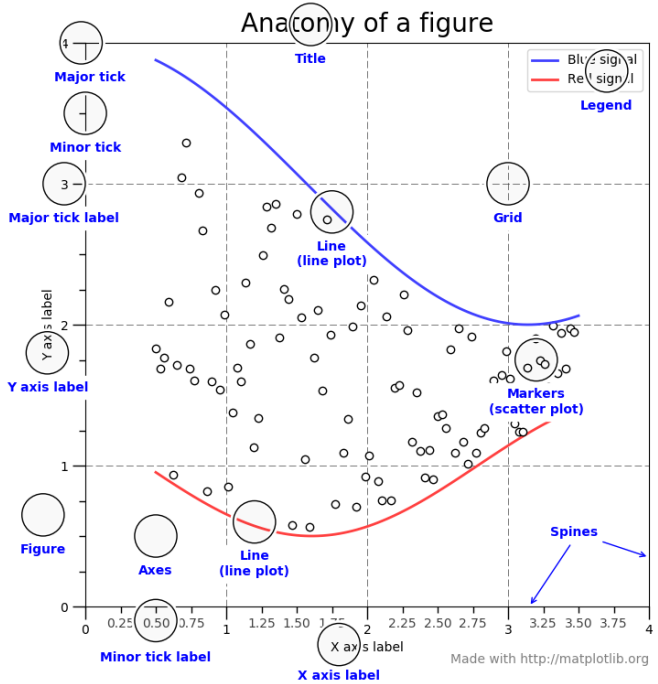
Dashboard with marketing KPIs
(Ad Spent, cost per thousand, ...)

Visualization tools: Plotting libraries (Python)

matplotlib

Advantages:
Widely used, many tools to customize plots (titles, legends, labels,...)

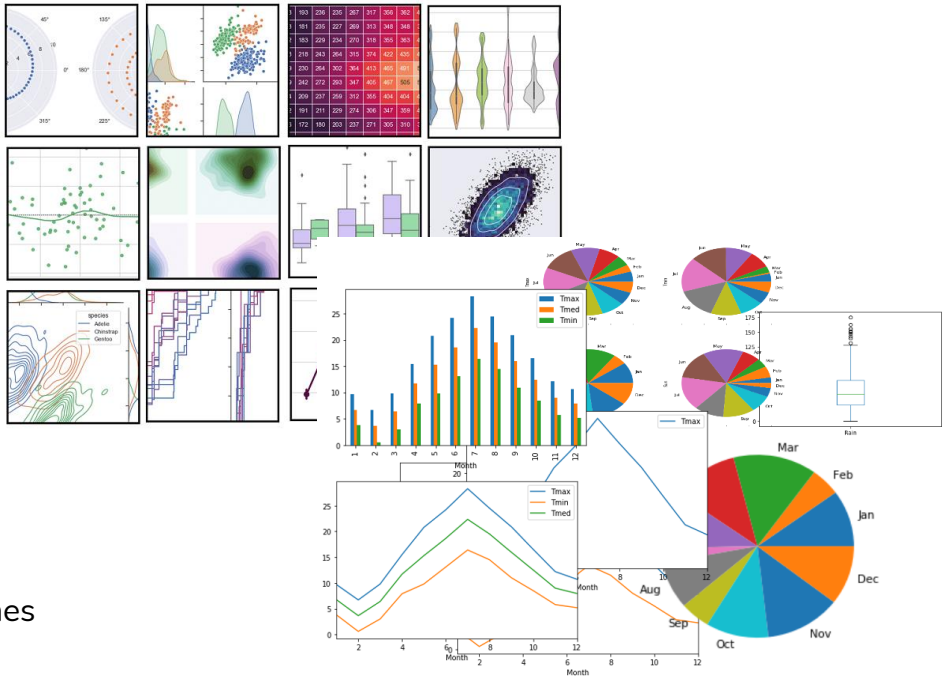
Disadvantages:
Verbose syntax, aesthetically lacking



Visualization tools: Plotting libraries (Python)

seaborn

- Less customizable but better aesthetics
- Many statistical plots & charts
- Better suited for datasets



pandas

- Easy API to create plots with Series/Dataframes
- Interface to Matplotlib plots

Visualization tools: Interactive plotting library

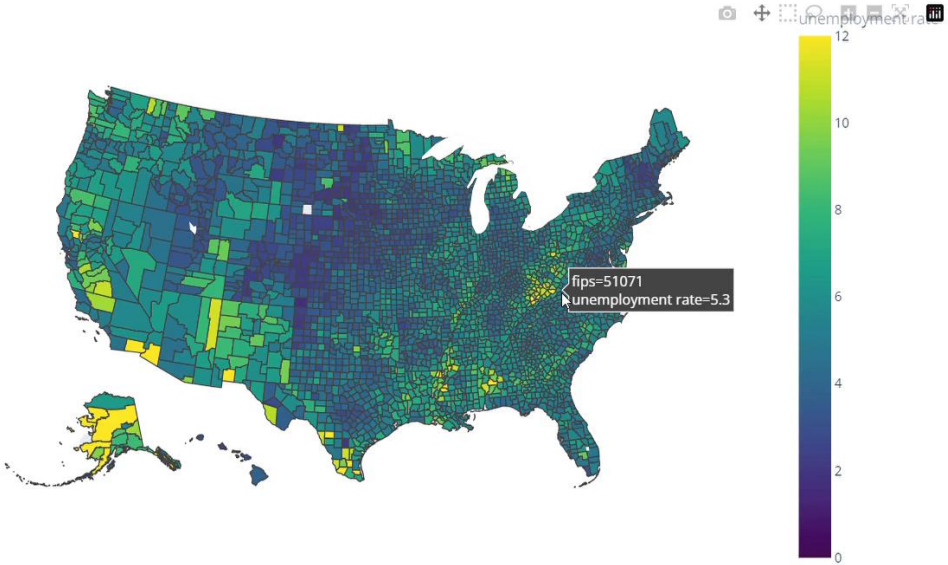
https://plotly.com/python/

plotly

Interactive plotting library
Create interactive plots & dashboards

Key features:

- Wide range of plots: charts, maps, ...
- High-level plotting library: **plotly express**
- Web application framework: **Dash**

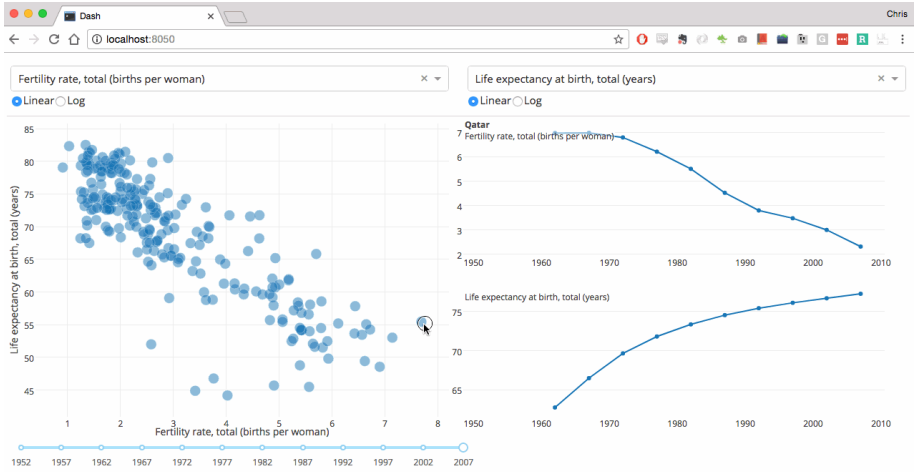


Interactive Choropleth map
Map of average unemployment rate
per US county (FIPS)

Visualization tools: Dashboard apps

plotly | Dash

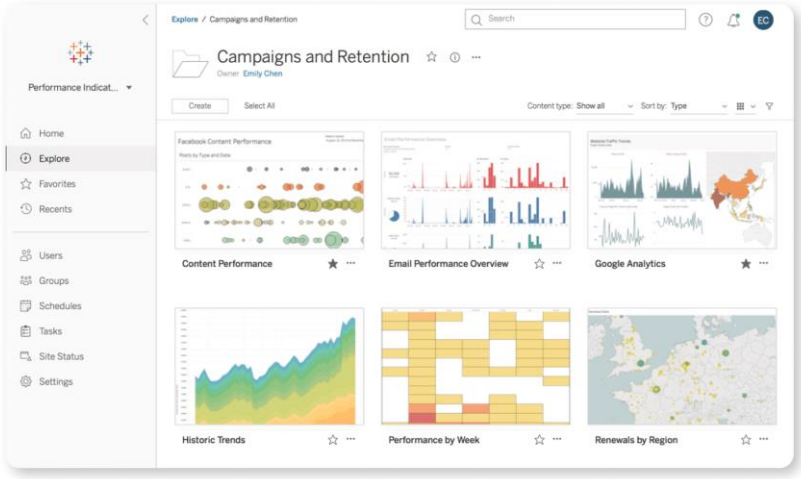
Plotly framework for web apps
Interactive dashboards with plotly components



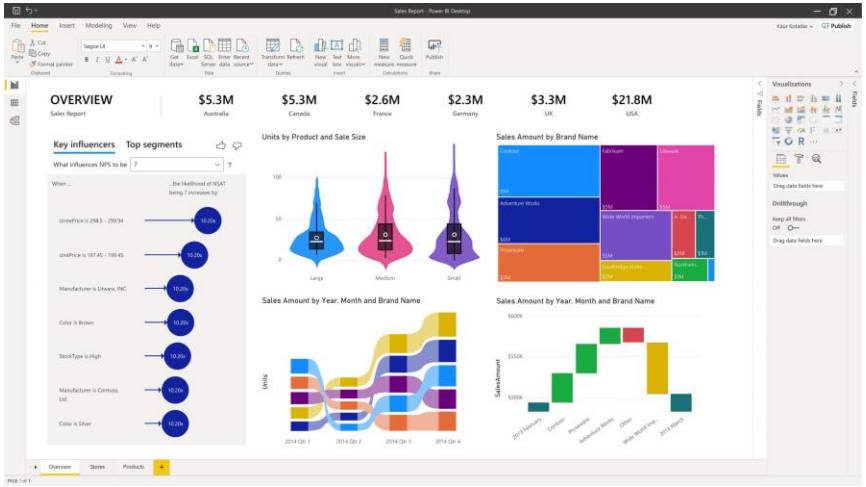
Key features:

- Build a web interface without full-stack development
- Deploy open-source apps with external hosting platforms
- Dash Entreprise: Create, share and deploy apps at scale

Visualization tools: Other frameworks



Tableau



Power BI

Business Analytics Softwares

Main types of visualization plots

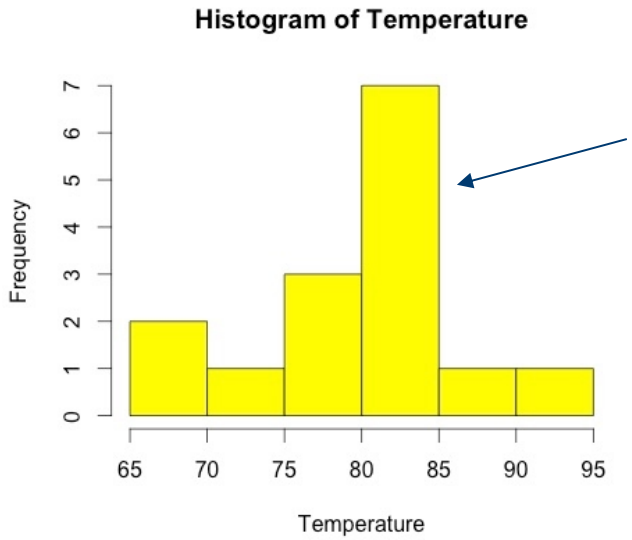
Distribution, Relationship, Comparison, Composition

Distribution plots: Histograms

- **Histograms** are graphs that represent the **distribution of a single variable**.
- They show **how frequently each range of values**, or "**bins**", appear in the dataset

Distribution plots: Histograms

- **Histograms** are graphs that represent the **distribution of a single variable**.
- They show **how frequently each range of values**, or "**bins**", appear in the dataset



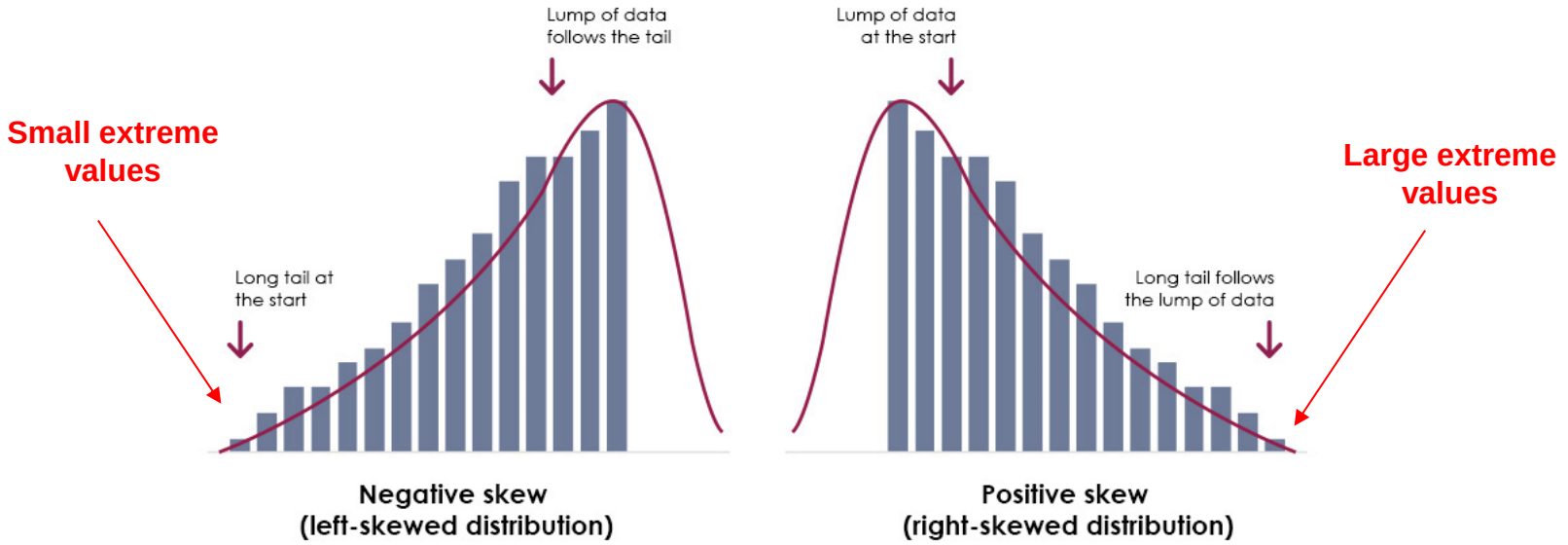
Each bin represents a **range of values** (65-70, 70-75, …)

Distribution plots: Histograms

- They are useful to study the **location**, **range** and **skewness** of a distribution
- They can help identify biases or extreme values in the data

Distribution plots: Histograms

- They are useful to study the **location**, **range** and **skewness** of a distribution
- They can help identify biases or extreme values in the data

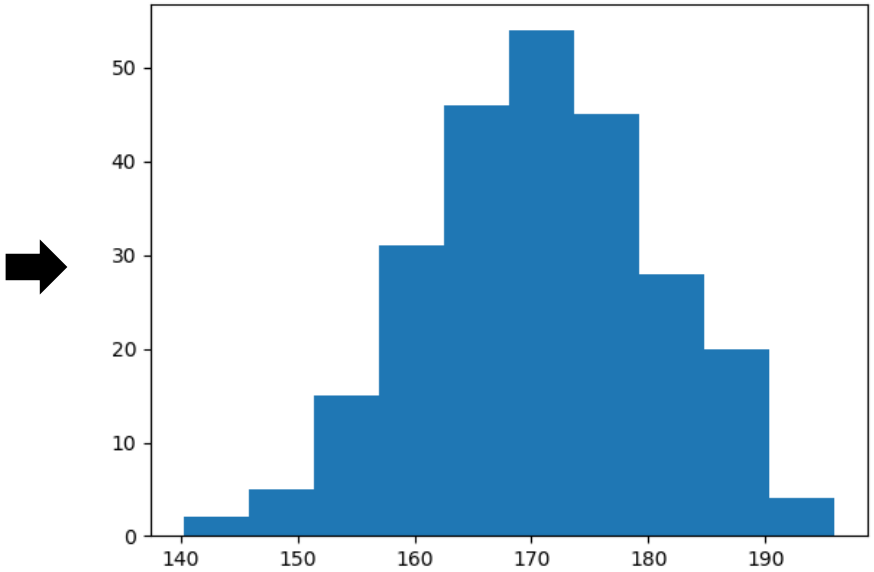


Histograms with Matplotlib

matplotlib

```
x = np.random.normal(170, 10, 250)

plt.hist(x)
plt.show()
```



Distribution plots: Boxplots

Boxplots are statistical plots that represent the distribution of a variable around its **median value**

Distribution plots: Boxplots

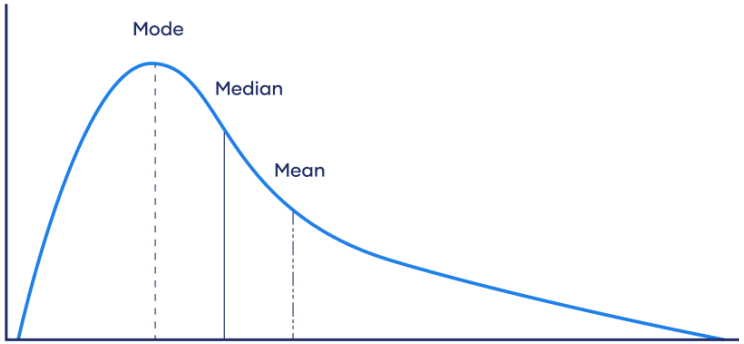
Boxplots are statistical plots that represent the distribution of a variable around its **median value**

What is the median value ?

The median divides the data into two equal halves

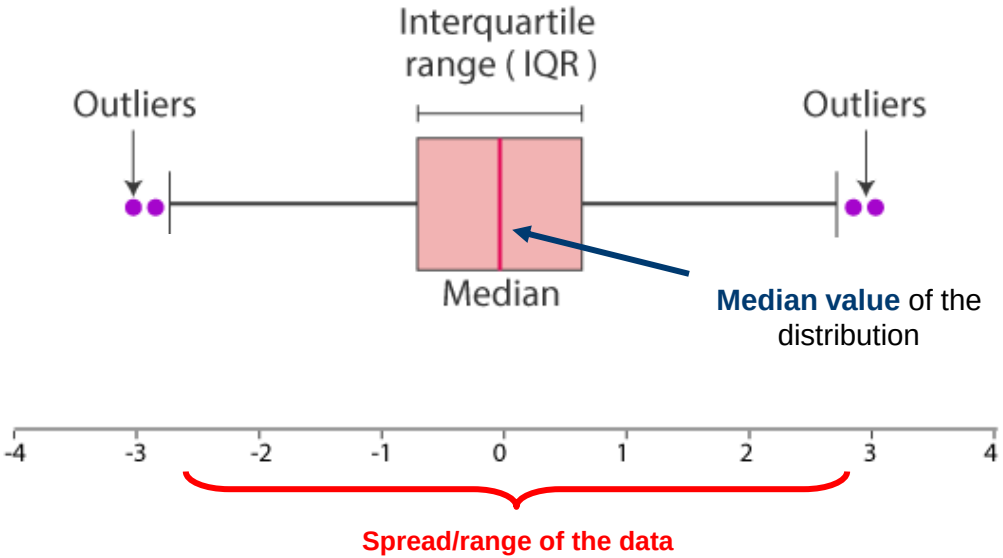
- 50% of values are less than or equal to it
- 50% are greater than or equal to it.

➡ *Isn't as sensitive to extreme values as the mean*



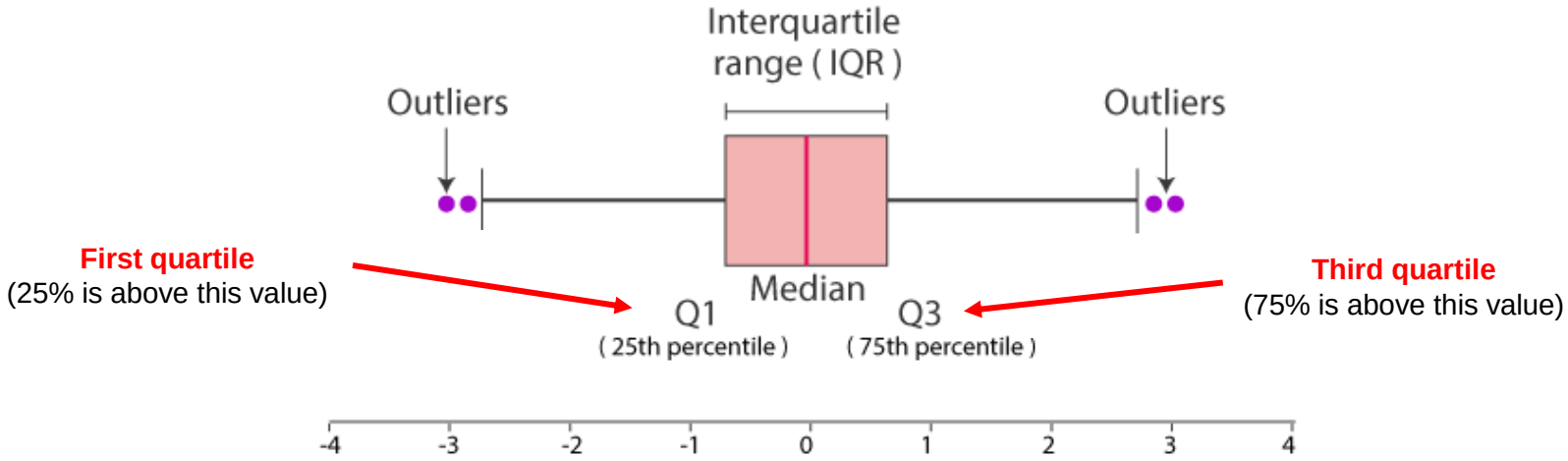
Distribution plots: Boxplots

Boxplots are statistical plots that represent the distribution of a variable around its **median value**



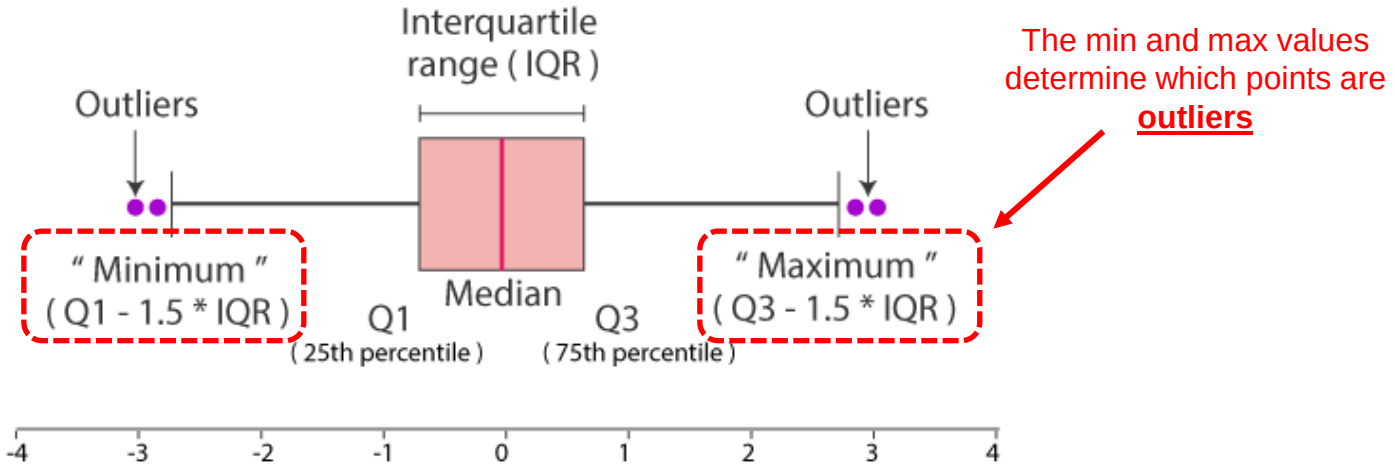
Distribution plots: Boxplots

Boxplots are statistical plots that represent the distribution of a variable around its **median value**



Distribution plots: Boxplots

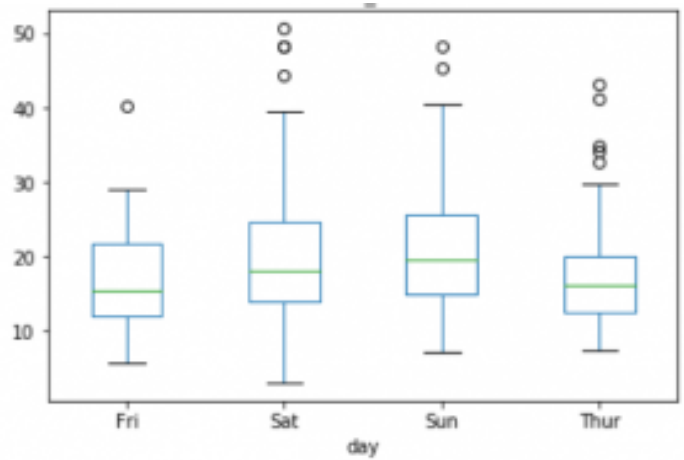
Boxplots are statistical plots that represent the distribution of a variable around its **median value**



Boxplots with Pandas

```
df.boxplot(by ='day', column =['total_bill'], grid = False)
```

	total_bill	tip	sex	smoker	day	time	size
0	16.99	1.01	Female	No	Sun	Dinner	2
1	10.34	1.66	Male	No	Sun	Dinner	3
2	21.01	3.50	Male	No	Sun	Dinner	3
3	23.68	3.31	Male	No	Sun	Dinner	2
4	24.59	3.61	Female	No	Sun	Dinner	4



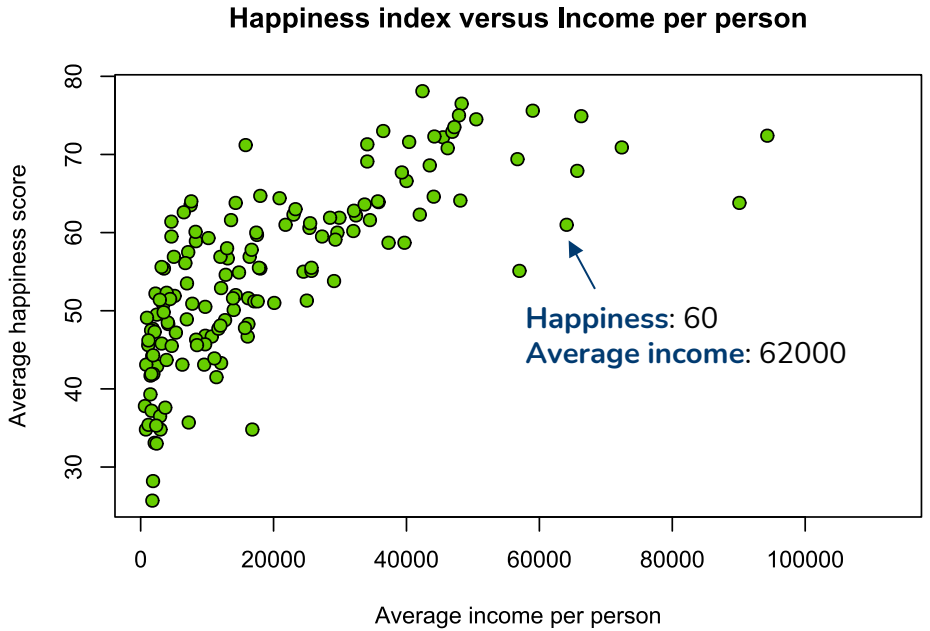
Relationship plots: Scatter plot

Scatter plots show the **relationship between two numerical variables** with dots (or any symbols)

Relationship plots: Scatter plot

Scatter plots show the **relationship between two numerical variables** with dots (or any symbols)

➡ *Each point represents a pair of values, one for each variable plotted*



Relationship plots: Scatter plot

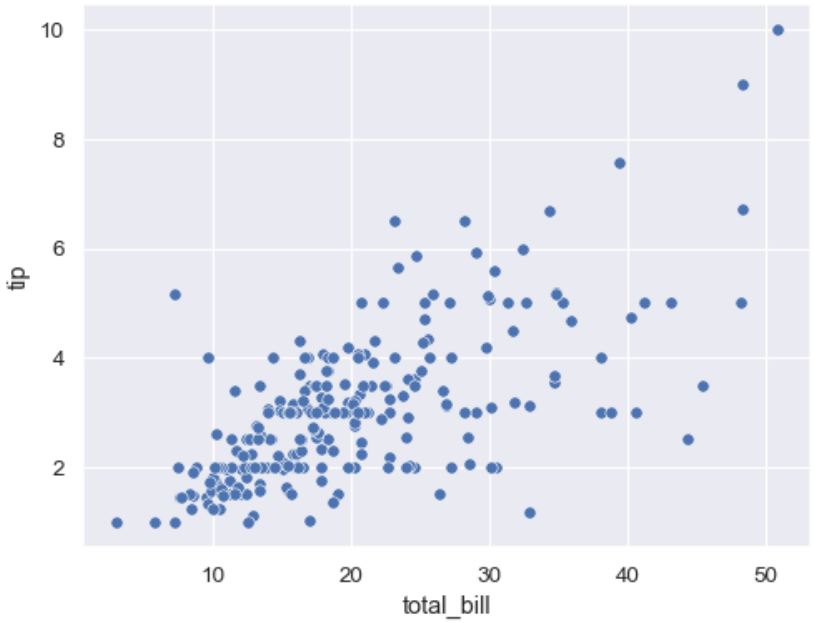
They can help identify **patterns within the data**, such as linear or non-linear relationships (or correlation).

Scatter plots with Seaborn

seaborn

```
sns.scatterplot(data=tips, x="total_bill", y="tip")
```

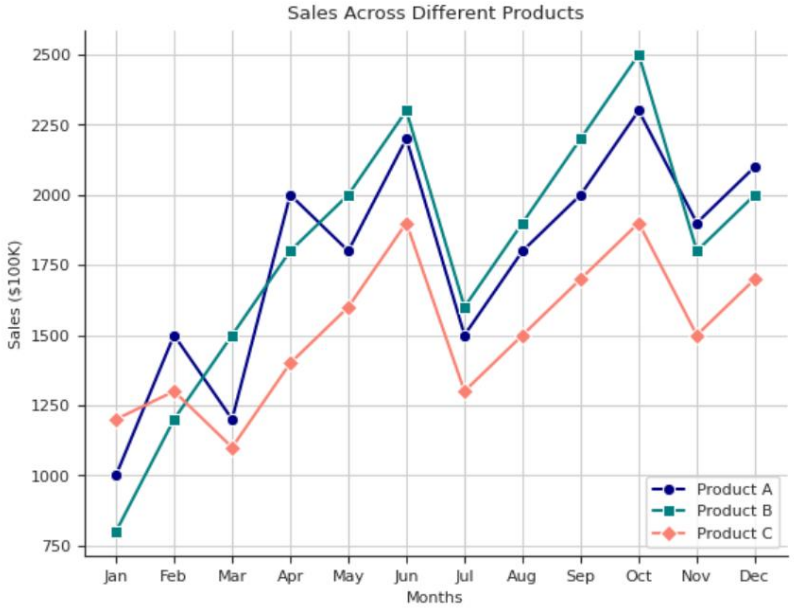
	total_bill	tip	sex	smoker	day	time	size
0	16.99	1.01	Female	No	Sun	Dinner	2
1	10.34	1.66	Male	No	Sun	Dinner	3
2	21.01	3.50	Male	No	Sun	Dinner	3
3	23.68	3.31	Male	No	Sun	Dinner	2
4	24.59	3.61	Female	No	Sun	Dinner	4



Comparison plots: Line plot

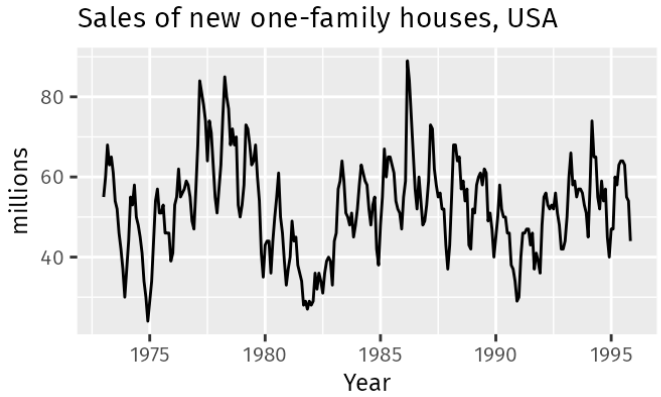
Line plots show the **evolution of numerical data** by building line segments between points

You can add multiple variables to a line plot using color

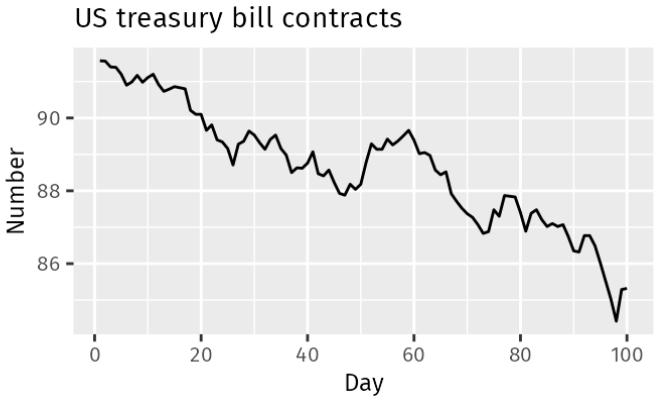


Comparison plots: Line plot

Line plots are used to identify **trends** and **seasonal patterns** in time series data (indexed by time)



Seasonal pattern

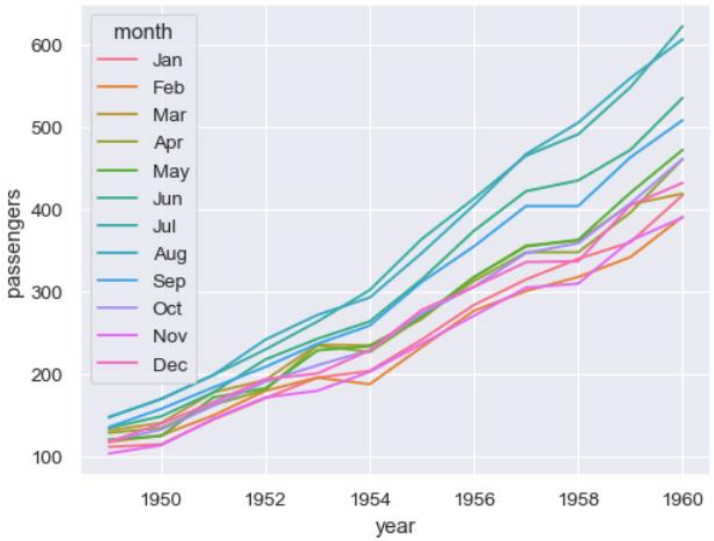


Trend

Line plot with Seaborn


```
sns.lineplot(data=flights, x="year", y="passengers", hue="month")
```

	year	month	passengers
0	1949	Jan	112
1	1949	Feb	118
2	1949	Mar	132
3	1949	Apr	129
4	1949	May	121



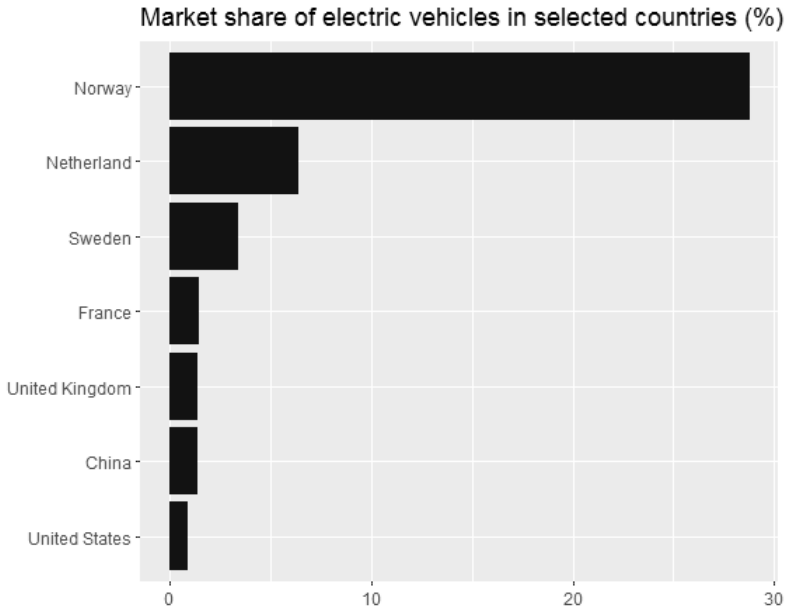
Comparison plots: Bar plot

Bar plots represent categorical data with rectangular bars. Each bar usually represents a potential value of the data.

Comparison plots: Bar plot

Bar plots represent categorical data with rectangular bars. Each bar usually represents a potential value of the data.

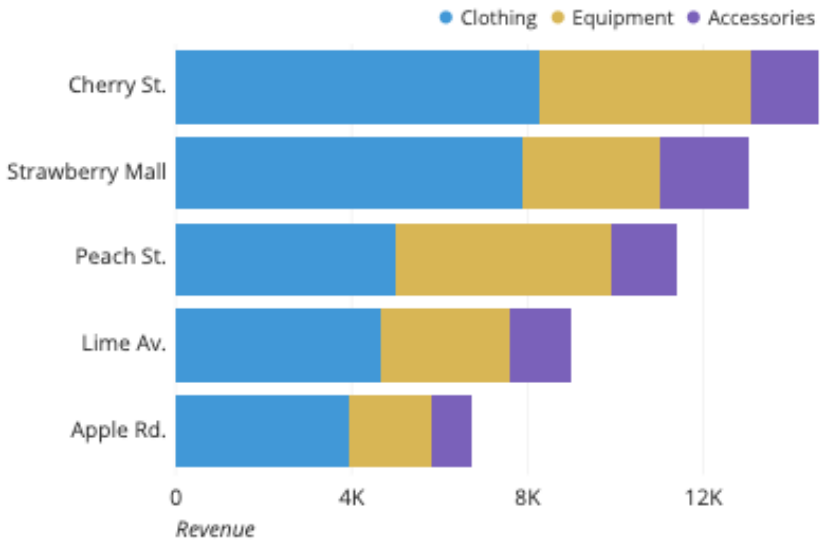
- Make **numerical comparisons** between different categories
- Bars can be represented **horizontally** or **vertically**



Comparison plots: Stacked bar plot

[Stacked bar plots] extend standard bar plot by looking at numerical values across two categorical variables

- Sub-bars (colors) represent the categories of the second variable



Revenue per store and product type

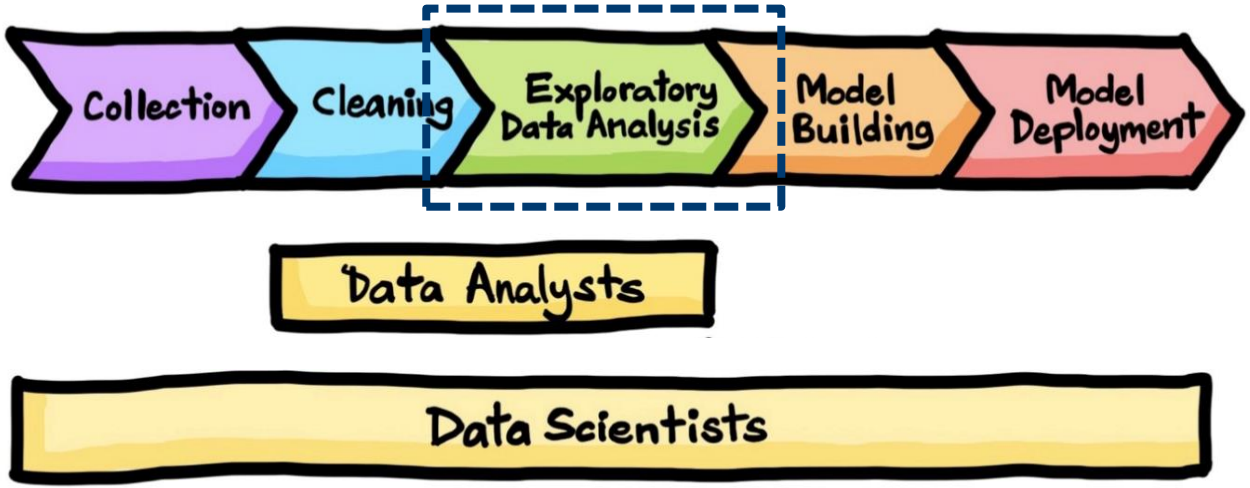
More plots: Python graph gallery

https://python-graph-gallery.com/

Exploratory Data Analysis (EDA)

What is Exploratory Data Analysis (EDA) ?

Exploratory Data Analysis is the process of performing initial investigations on data with the help of **summary statistics** and **graphical representations**



What is Exploratory Data Analysis (EDA) ?

Exploratory Data Analysis is the process of performing initial investigations on data with the help of **summary statistics** and **graphical representations**

Why is it useful ?

- Understand the data better
- Uncover **underlying patterns** or **biases**
- Detect **anomalies** or extreme values
- Check assumptions before building a predictive model

What is Exploratory Data Analysis (EDA) ?

We will use the **titanic prediction** dataset as an example for Exploratory Data Analysis (EDA).

➡ Goal is to predict whether a **passenger survived the titanic**

Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare
0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500
1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833
1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250
1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000

Types of Exploratory Data Analysis

Univariate Analysis

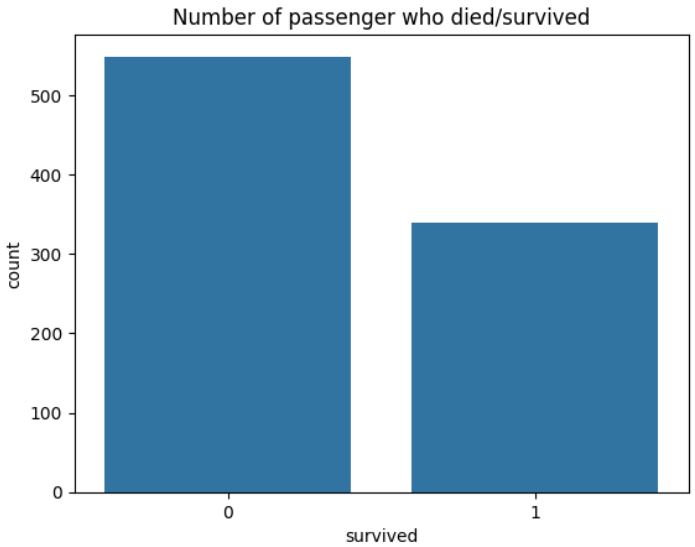
Study the characteristics of the dataset's variables individually

Examples:

- Number of missing values
- Data types (int/float, object, …)
- Frequency of possible values (categorical data)
- Summary statistics (continuous data)

Types of Exploratory Data Analysis

Univariate Analysis



Frequency of each category

Get the number of passenger that died and survived

Number of non-missing values

Pandas dtypes

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 12 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   PassengerId  891 non-null    int64
 1   Survived     891 non-null    int64
 2   Pclass       891 non-null    int64
 3   Name         891 non-null    object
 4   Sex          891 non-null    object
 5   Age          714 non-null    float64
 6   SibSp        891 non-null    int64
 7   Parch        891 non-null    int64
 8   Ticket       891 non-null    object
 9   Fare         891 non-null    float64
 10  Cabin        204 non-null    object
 11  Embarked     889 non-null    object
dtypes: float64(2), int64(5), object(5)
memory usage: 83.7+ KB
```

Missing values & data types

Get general information on the data

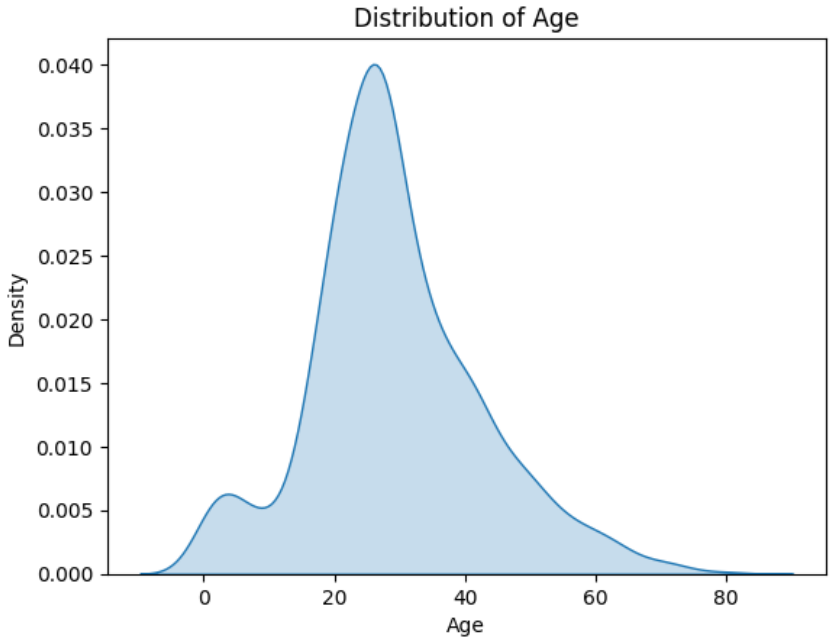
Types of Exploratory Data Analysis

Univariate Analysis

	Age	Fare
count	889.000000	889.000000
mean	29.336524	32.096681
std	13.226753	49.697504
min	0.420000	0.000000
25%	22.000000	7.895800
50%	27.000000	14.454200
75%	36.000000	31.000000
max	80.000000	512.329200

Statistical analysis

Compute summary statistics for Age and Fare



Individual distribution

Study the range skewness of ages in the data

Types of Exploratory Data Analysis

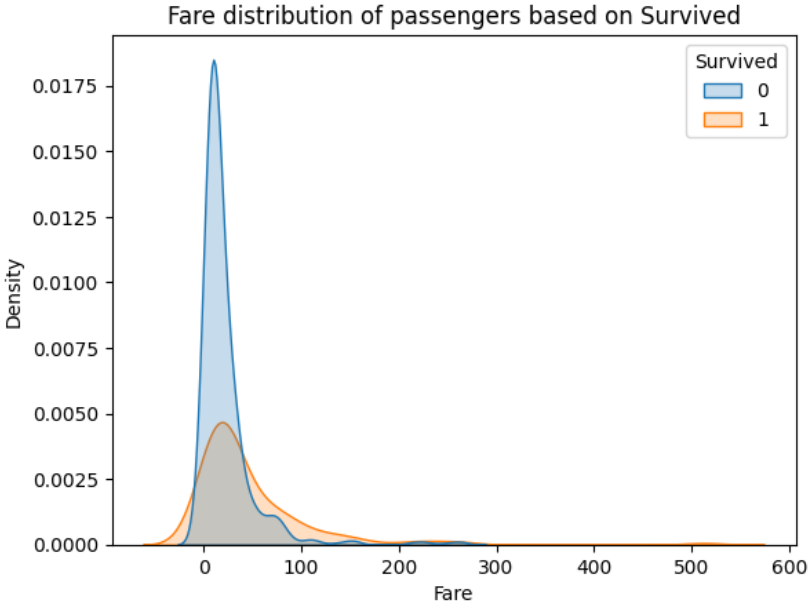
Bivariate Analysis

Analyze the relationship between two variables in the dataset

- Detect **unknown patterns/trends** between two variables
- Understand how a **variable could explain the variable you are trying to predict** (for a predictive analysis)

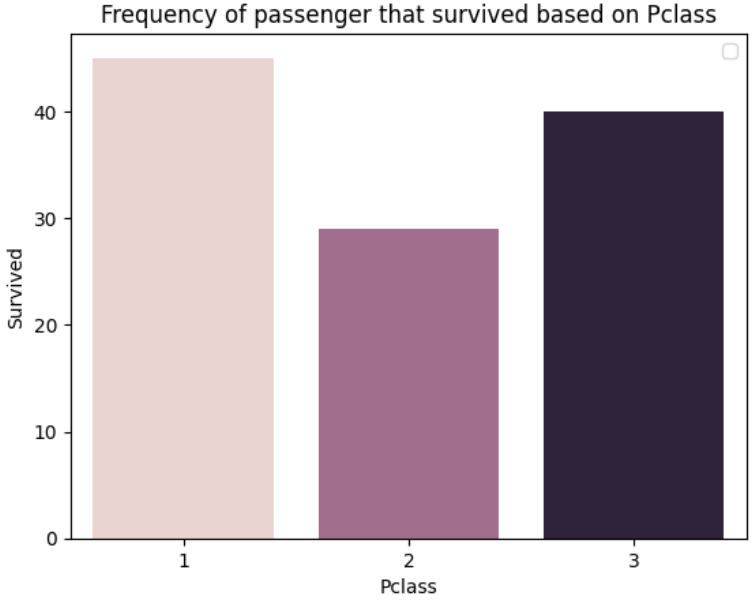
Types of Exploratory Data Analysis

Bivariate Analysis



1 continuous and 1 categorical

Split the Fare distribution based on the "Survived" values

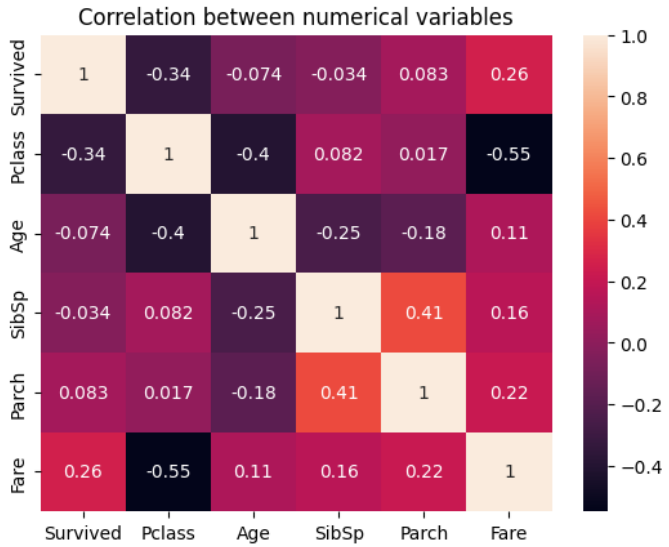


2 categorical variables

Plot the number of passengers that "Survived" based on the passenger's class

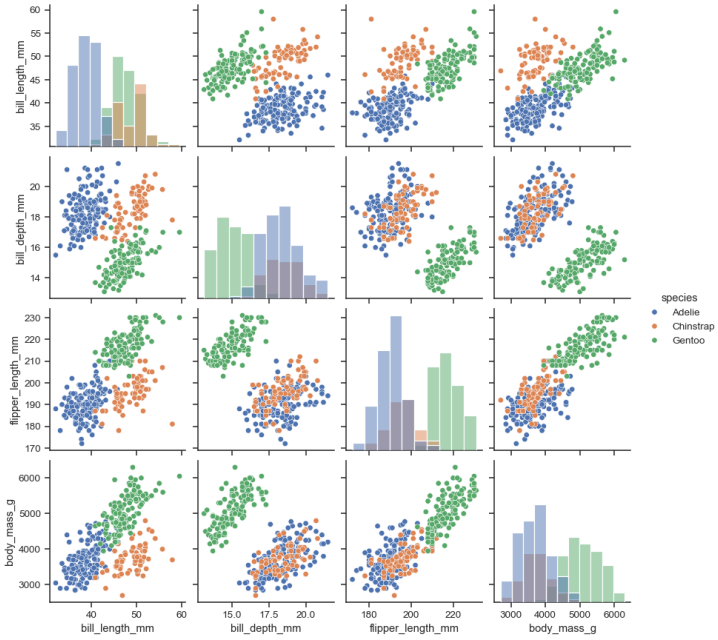
Types of Exploratory Data Analysis

Bivariate Analysis



Correlation analysis

Detect *redundant variables* or *strong correlations* with the variable to predict



Pair plot

Build bivariate scatter plots for each continuous variable

Thank you for listening ! 🤗

Do you have any questions ?

HEC PARIS

INSTITUT POLYTECHNIQUE DE PARIS

Hi! PARIS

PARIS ARTIFICIAL INTELLIGENCE FOR SOCIETY