Institute for advanced architecture of Catalonia | BARCELONA

MASTER IN CITY & TECHNOLOGY
DIGITAL TOOLS AND BIG DATA
2019/2020

FACULTY DIEGO PAJARITO

Data ingestion

Getting back to pandas

Pycharm
Python
Github
Anaconda
QGIS

Libraries
- Pandas
- Seaborn
- Matplotlib
- Numpy
- geopandas

PANDAS - Python data analysis library

Data analysis tools for the Python programming language
Open source, BSD-licensed library
High-performance, easy-to-use data structures

Using Conda

```
conda install pandas
```

Using PIP

```
python -m pip install --upgrade pandas
```

Using Conda's graphical interface or Pycharm

Pandas documentation available here: https://pandas.pydata.org/

List:	Dictionary:	Tuple:

['Tokyo',
'Delhi',
'Shanghai',
'Sao Paulo',
'Mexico City',
'Cairo',
'Dhaka',
'Mumbai',
'Beijing',
'Osaka']
	{
	'cities': ['Tokyo','Delhi', 'Shanghai',
	'Sao Paulo', 'Mexico City', 'Cairo',
	'Dhaka', 'Mumbai', 'Beijing', 'Osaka'],
	'population': (37435191, 29399141, 26317104,
	21846507, 21671908, 20484965, 20283552,
	20185064, 20035455, 19222665),
	'source':
	'http://worldpopulationreview.com/world-cities/'
	}
		(37435191,
		29399141,
		26317104,
		21846507,
		21671908,
		20484965,
		20283552,
		20185064,
		20035455,
		19222665)

if/else/elif:	For/While:

`if a == 0:`
`  print('A equals 0')`
`else:`
`  Print('A not 0')`
	`for i in range(10):`
	`  print('i equals: ' + str(i))`

id	name	address	postal_code	lon	lat
1	IAAC main building	Carrer de Pujades 102	08005	2.1932315826416016	41.395747068298895
2	IAAC atelier	Carrer de Pujades 59	08005	2.1919387578964233	41.39522593585012

Get Column / Columns / Row / Rows	Get Subset

`df['column']`
`df[['column1', 'column2']]`
`df[0:1]`
`...`
	`df.head()`
	`df.iloc[ : , 2 ]`
	`df.loc()`
	`...`

Try to create your own dataframe, add some values and see how you can create it from lists, tuples or dicts.

Variables & Statistics

The ABC of data analytics

Variables can take multiple values that change across the time.

We use variables to describe objects, so they need a clear definition.

Some elements to consider when defining variables are:

Name
Description
Purpose
Domain

And, therefore, data structure

A key element that defines data structure is the Domain.

It refers to the kind of values a variable can take and define the most convenient data type to use.

Boolean	True / False
Numeric	Int: …, -3, -2, -1, 0, 1, 2, 3, ... Float: -∞, …, 0.0, …, ∞
Char	‘a’ ‘word’ ‘something else !”·$%&(/’ ‘../meaningful/text/file.ext’
Objects	Lists [ ] Tuple () Dict {‘key’: ‘value} List of lists [[ ], [ ]] Dataframe

When having a large or medium data set there is a need to summarise the information such a set provides. Tendencies and dispersion are two relevant features to identify.

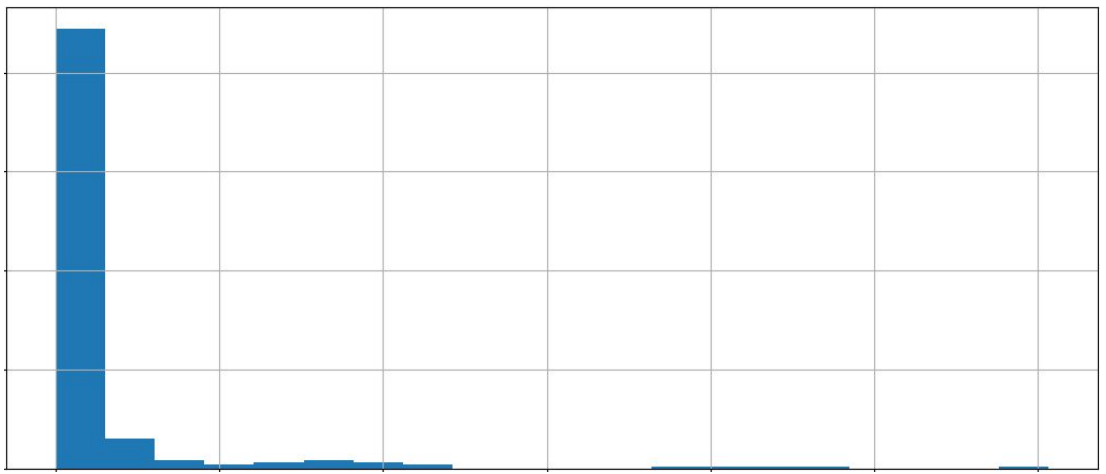
What can we say about the place in which we are located?

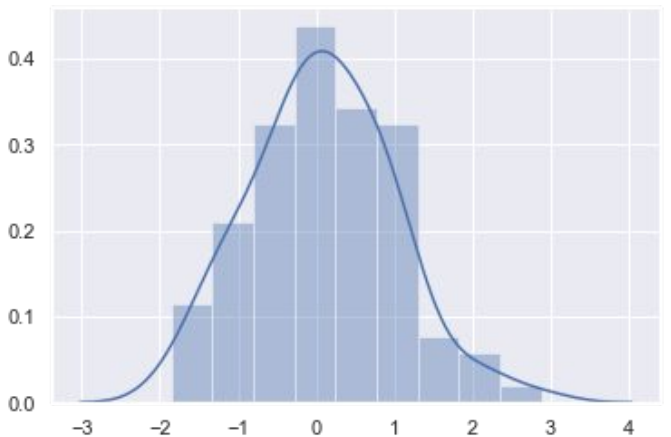
Central tendency	Mean Mode Median Frequency (integer, classes)
Dispersion	Range Variance Standard Deviation

Density & Frequency

The ABC of data analytics

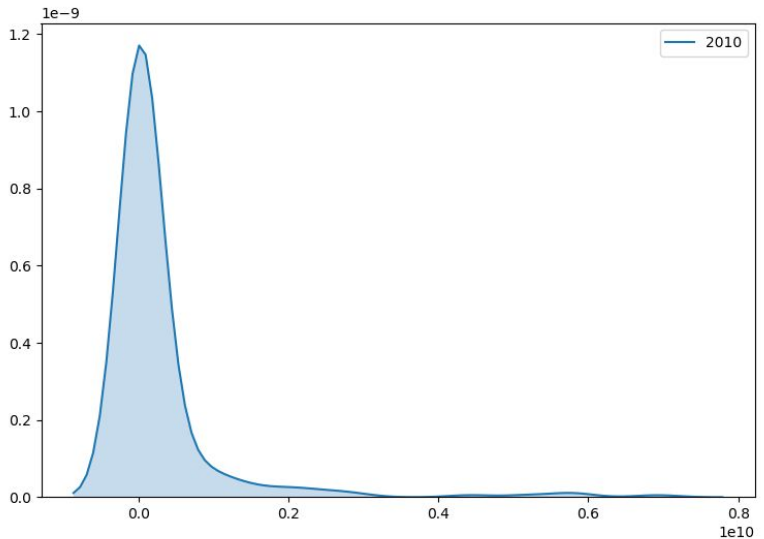
We can use either matplotlib or seaborn to create the histograms. **Histograms** are representations of the distribution of numeric data. After establishing ranges, the numeric values **falling into these** ranges and counted and represented in a plot. The equivalent diagram for non-numeric data is based on **frequency tables**.





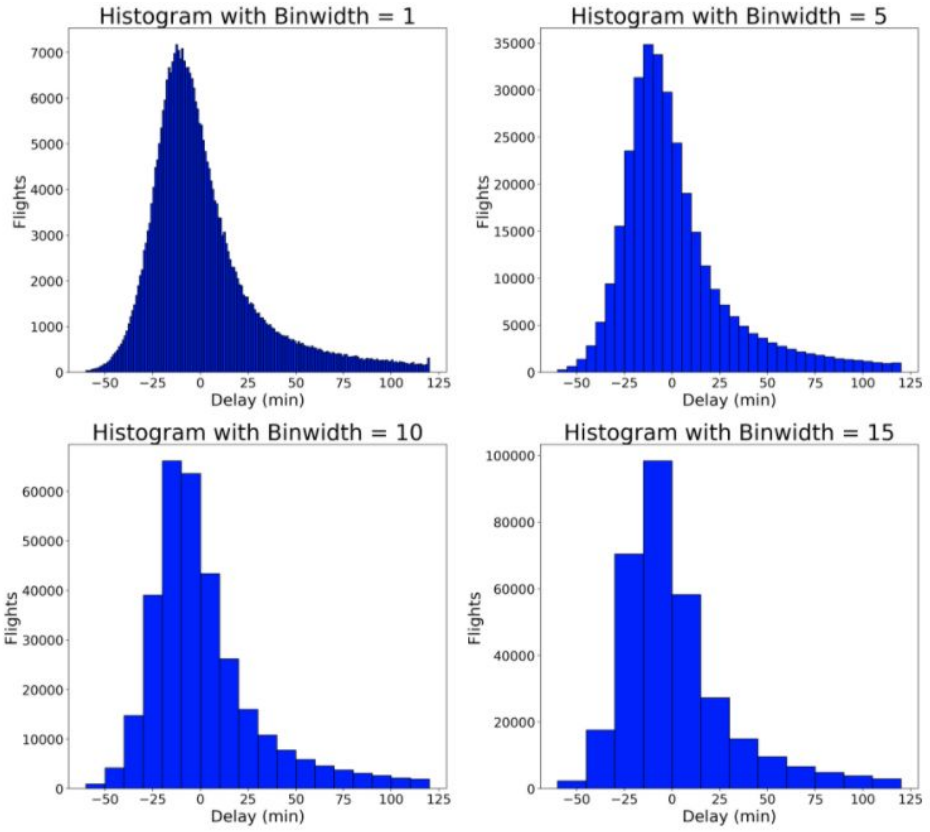
Normal distribution is a common reference for numeric distribution. It assumes a random process to generate data that is usually not the case. It is built from the mean and standard deviation values that can be overlayed to the histogram.

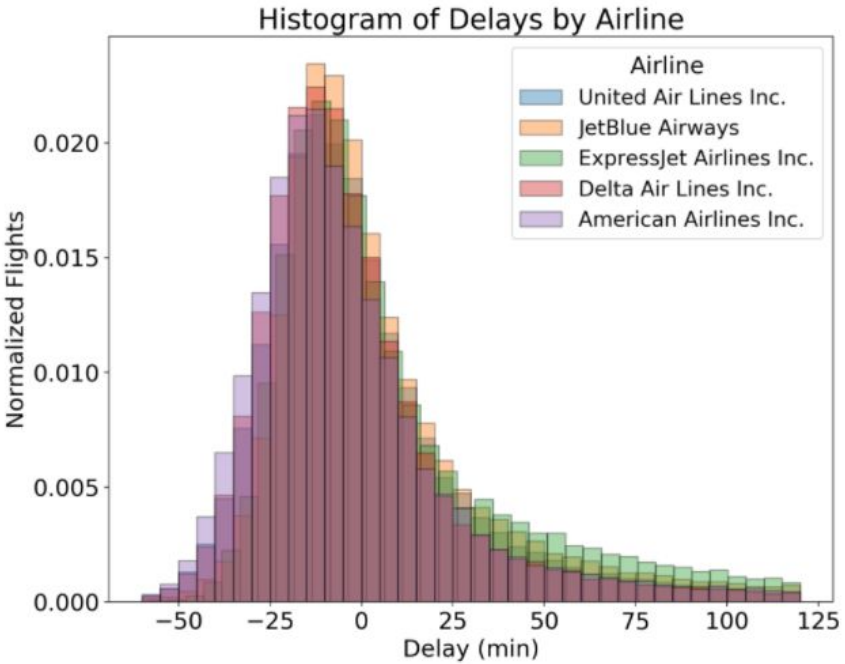
We can use either matplotlib or seaborn to create the histograms, probability distribution and density charts.

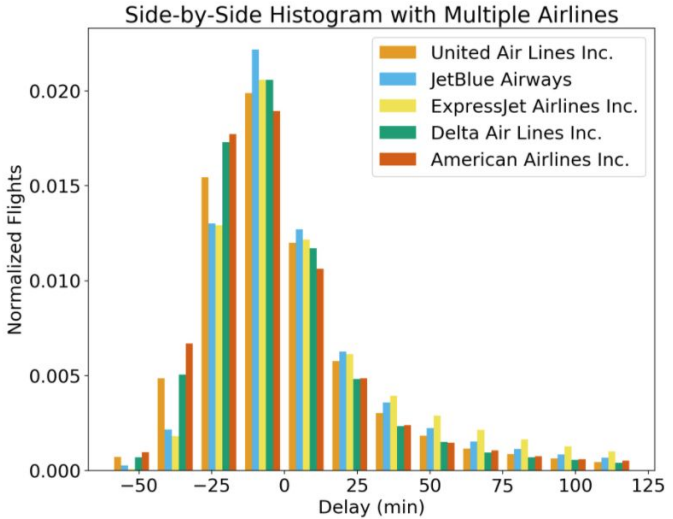


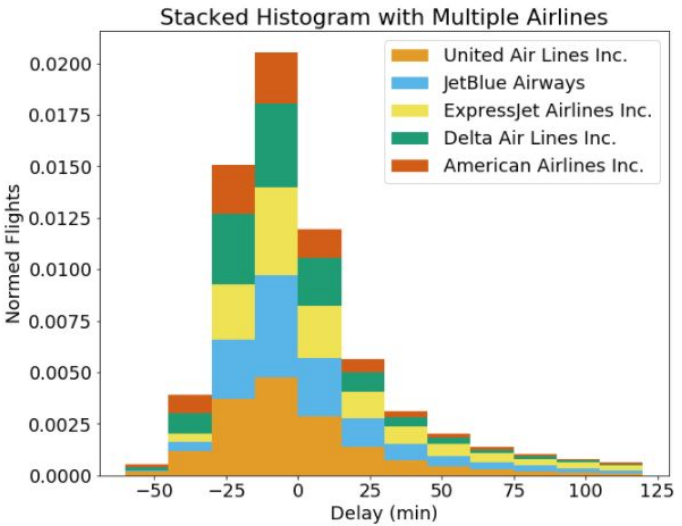
This [blog post] summarizes the relevance of identifying and describing frequency.

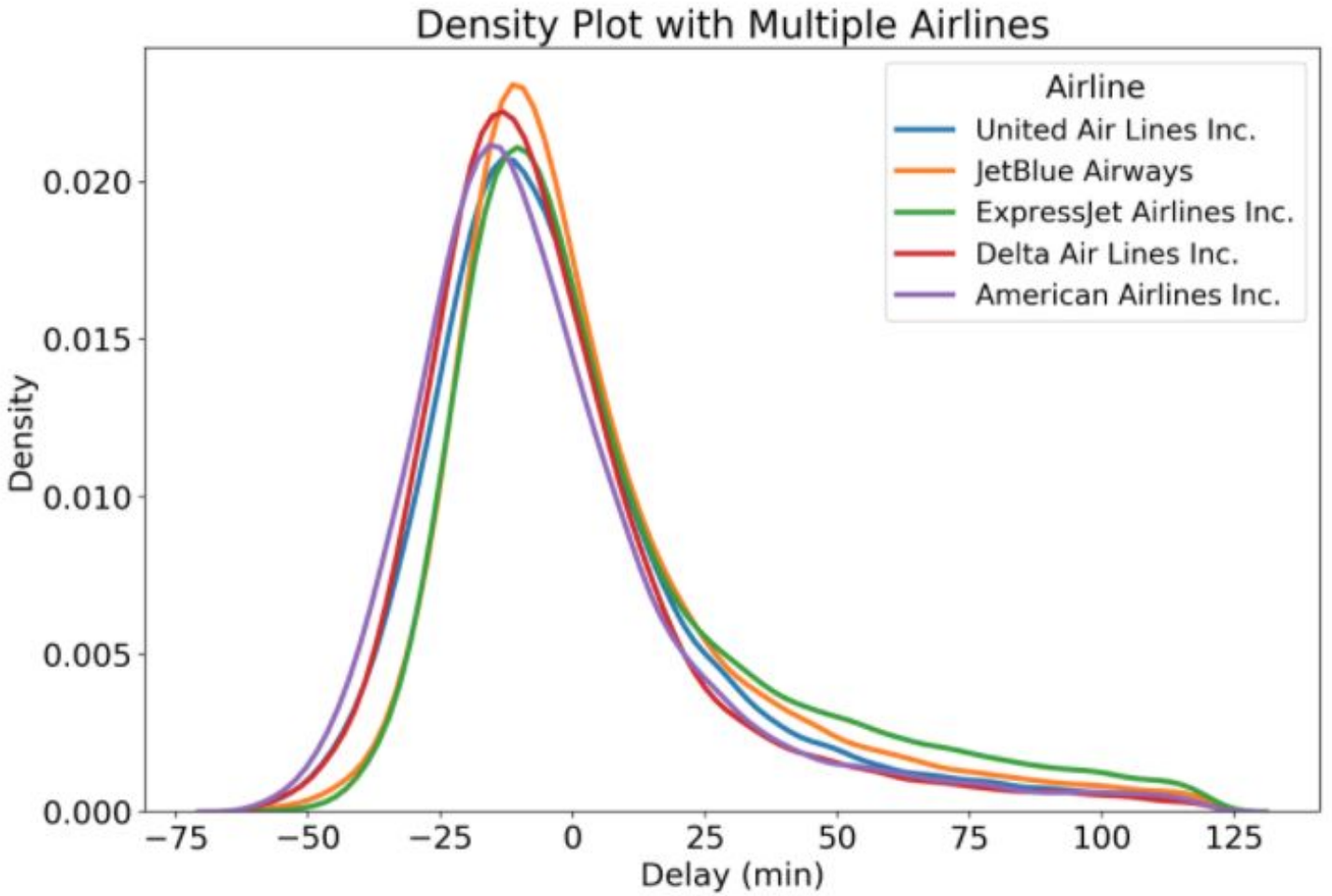
	arr_delay	name
0	11.0	United Air Lines Inc.
1	20.0	United Air Lines Inc.
2	33.0	American Airlines Inc.
3	-18.0	JetBlue Airways
4	-25.0	Delta Air Lines Inc.
5	12.0	United Air Lines Inc.
6	19.0	JetBlue Airways
7	-14.0	ExpressJet Airlines Inc.
8	-8.0	JetBlue Airways
9	8.0	American Airlines Inc.




Histogram of Delays by Airline
Airline
United Air Lines Inc.
JetBlue Airways
ExpressJet Airlines Inc.
Delta Air Lines Inc.
American Airlines Inc.
Normalized Flights
0.020
0.015
0.010
0.005
0.000
−50
−25
0
25
50
75
100
125
Delay (min)


Side-by-Side Histogram with Multiple Airlines
United Air Lines Inc.
JetBlue Airways
ExpressJet Airlines Inc.
Delta Air Lines Inc.
American Airlines Inc.
Normalized Flights
0.020
0.015
0.010
0.005
0.000
−50
−25
0
25
50
75
100
125
Delay (min)

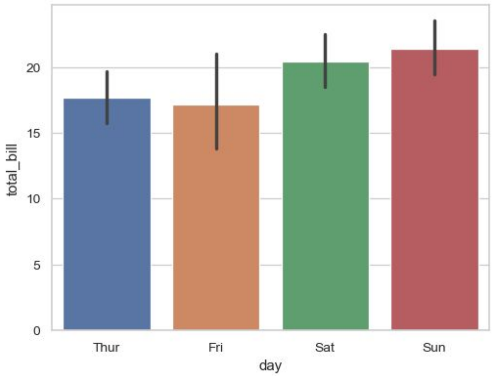

Stacked Histogram with Multiple Airlines
United Air Lines Inc.
JetBlue Airways
ExpressJet Airlines Inc.
Delta Air Lines Inc.
American Airlines Inc.
Normed Flights
0.0200
0.0175
0.0150
0.0125
0.0100
0.0075
0.0050
0.0025
0.0000
−50
−25
0
25
50
75
100
125
Delay (min)

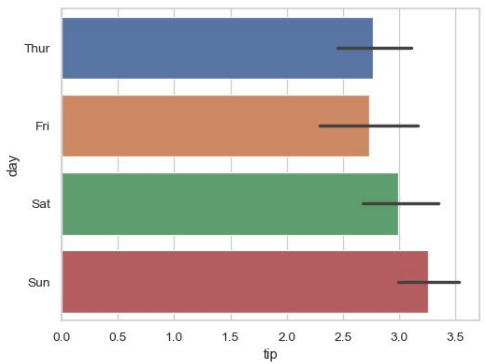

Density Plot with Multiple Airlines
Airline
United Air Lines Inc.
JetBlue Airways
ExpressJet Airlines Inc.
Delta Air Lines Inc.
American Airlines Inc.
Density
0.020
0.015
0.010
0.005
0.000
−75
−50
−25
0
25
50
75
100
125
Delay (min)

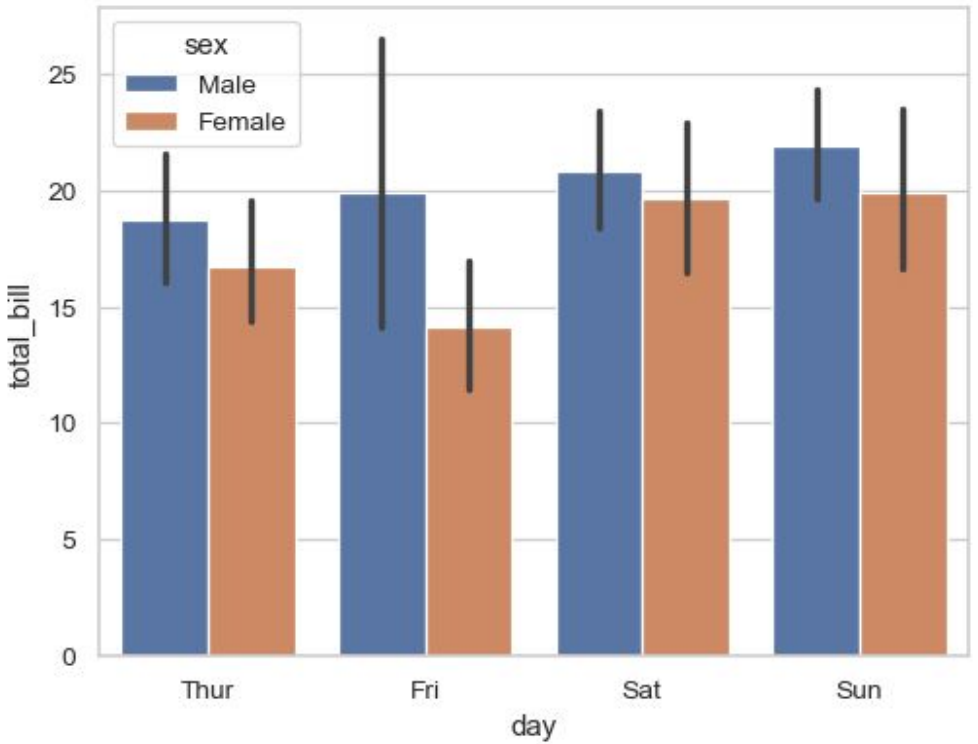
Bar plots

The ABC of data analytics

Bar plots include 0 in the quantitative axis range, and they are a good choice when 0 is a meaningful value for the quantitative variable, and you want to make comparisons against it.







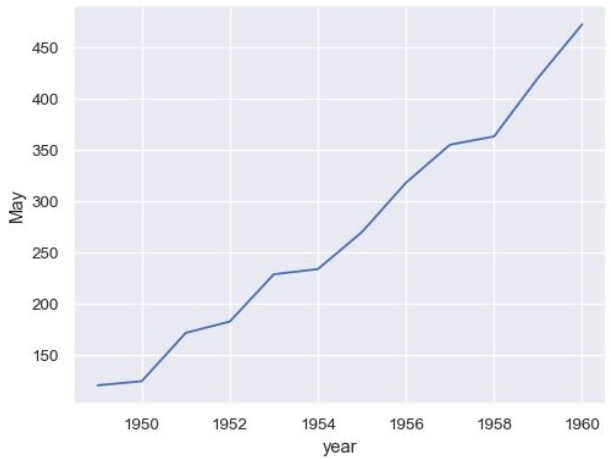
Source: https://seaborn.pydata.org/generated/seaborn.barplot.html

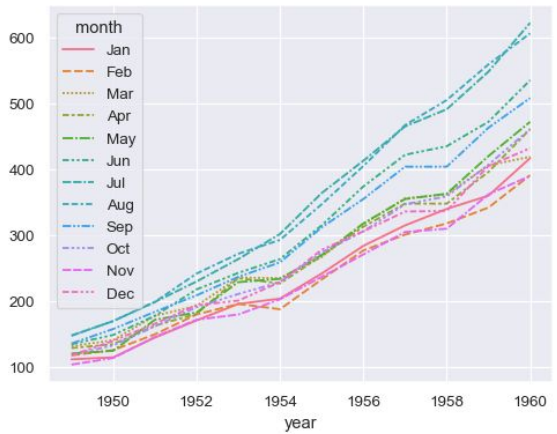
Line plots

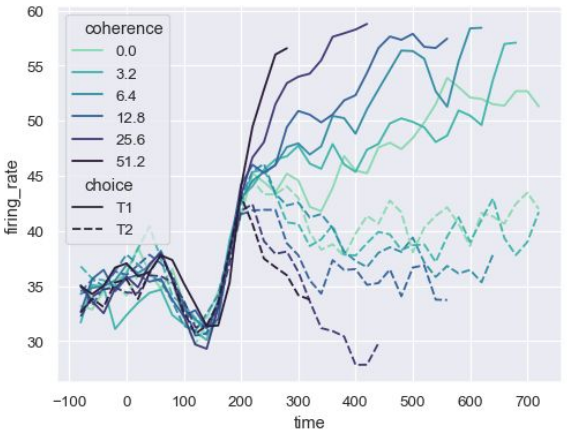
The ABC of data analytics

Hue, size, and style are parameters that serve to control what visual semantics are used to identify the different subsets. It is possible to show up to **three** dimensions independently by using all three semantic types, but this style of plot can be hard to interpret and is often ineffective.

Using redundant semantics (i.e. both hue and style for the same variable) can be helpful for making graphics more accessible.





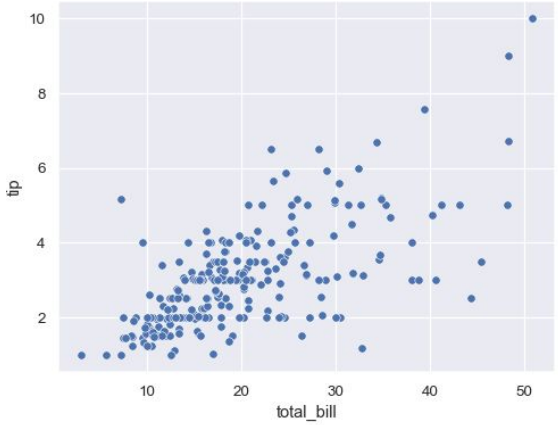


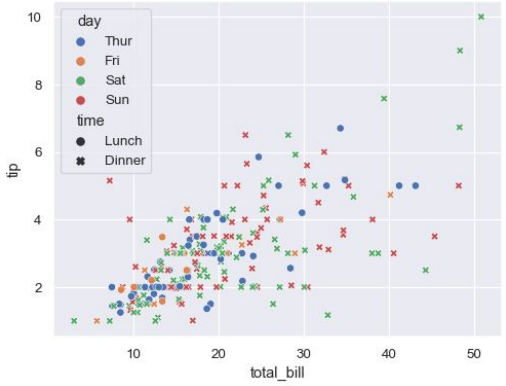
Source: https://seaborn.pydata.org/generated/seaborn.lineplot.html

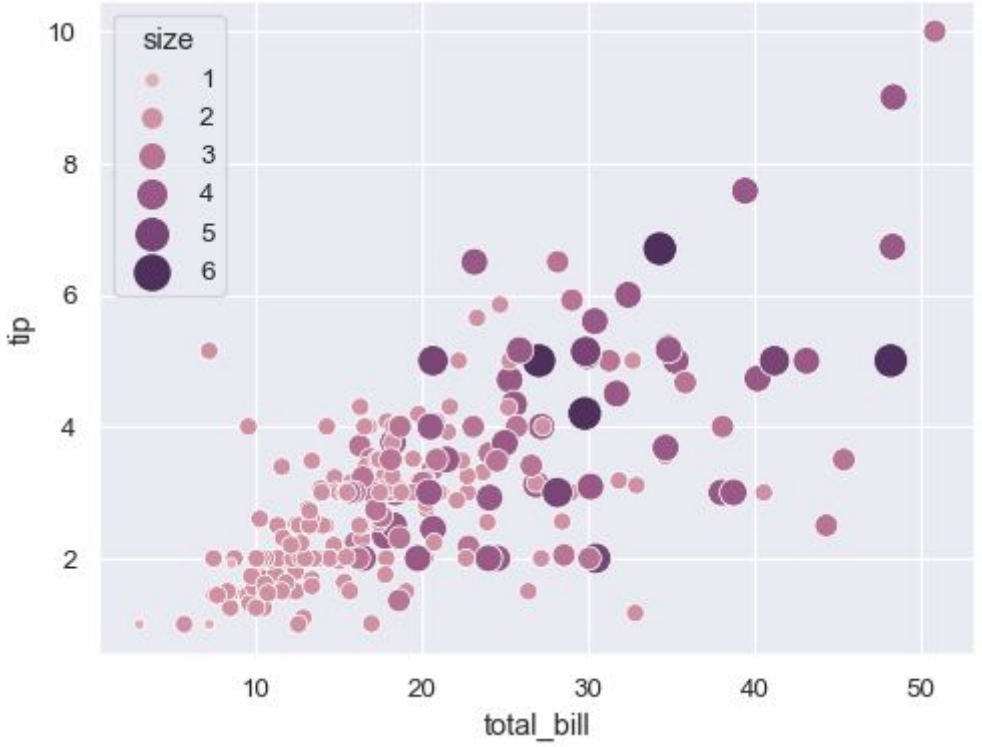
Scatter plots

The ABC of data analytics

Using redundant semantics (i.e. both hue and style for the same variable) can be helpful for making graphics more accessible.







Source: https://seaborn.pydata.org/generated/seaborn.scatterplot.html

To go beyond in Pandas

https://pandas.pydata.org/pandas-docs/stable/getting_started/10min.html#min

https://matplotlib.org/3.1.1/users/index.html

https://towardsdatascience.com/matplotlib-tutorial-learn-basics-of-pythons-powerful-plotting-library-b5d1b8f67596

https://seaborn.pydata.org/tutorial.html

https://elitedatascience.com/python-seaborn-tutorial

And QGIS

https://docs.qgis.org/3.4/en/docs/training_manual/index.html

Just to feed your curiosity:

Bivariate and pairwise plots as well as Spatial Clustering and geometry generator

Institute for advanced architecture of Catalonia | BARCELONA

MASTER IN CITY & TECHNOLOGY
DIGITAL TOOLS AND BIG DATA
2019/2020

FACULTY DIEGO PAJARITO