

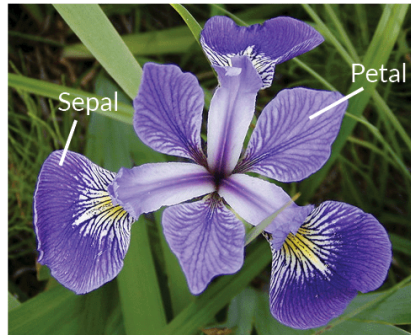
Introduction to dataframes



QB Bootcamp, Day 2
Thursday, 29 August 2019
1:00pm - 1:30pm

What are dataframes?

Fisher's "Iris" dataset is a famous example dataset in statistics and dataviz



Iris Versicolor



Iris Setosa



Iris Virginica

R. A. Fisher (1936).
"The use of multiple measurements
in taxonomic problems".
Annals of Eugenics. 7 (2): 179–188.
(data collected by Edgar Anderson)

150 rows
(50 per species)

	sepal_length	sepal_width	petal_length	petal_width	species
0	5.1	3.5	1.4	0.2	setosa
1	4.9	3.0	1.4	0.2	setosa
2	4.7	3.2	1.3	0.2	setosa
3	4.6	3.1	1.5	0.2	setosa
4	5.0	3.6	1.4	0.2	setosa

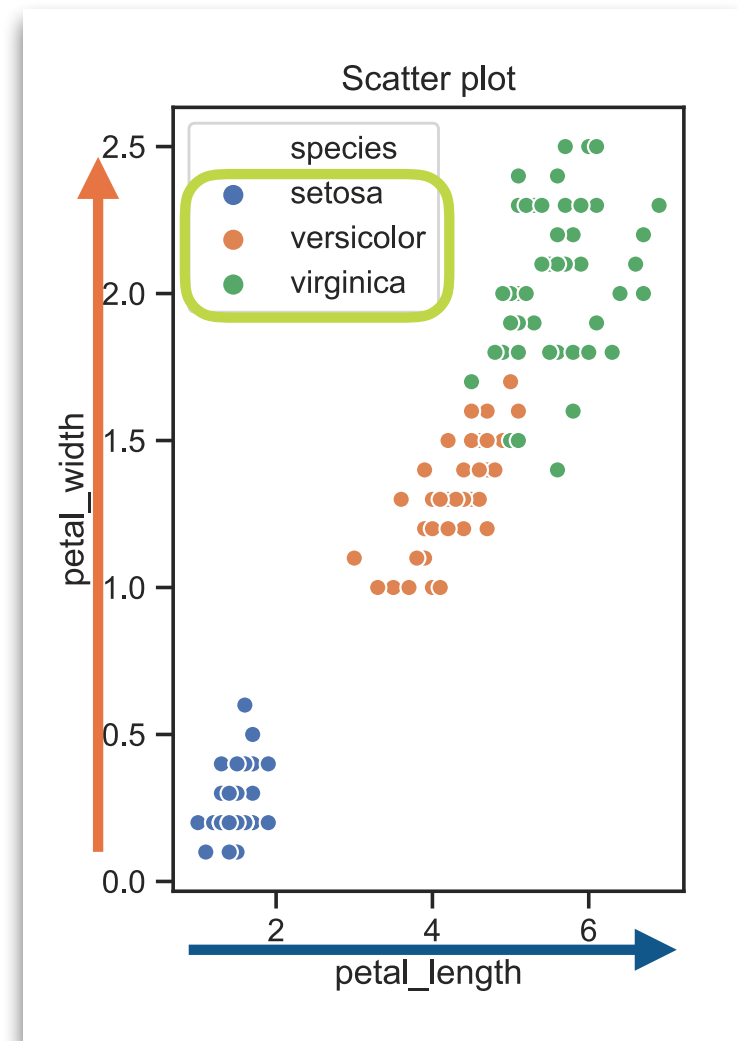
⋮

Why use dataframes?

Dataframes greatly facilitate data visualization

“aesthetics”

	sepal_length	sepal_width	x petal_length	y petal_width	color species
0	5.1	3.5	1.4	0.2	setosa
1	4.9	3.0	1.4	0.2	setosa
2	4.7	3.2	1.3	0.2	setosa
3	4.6	3.1	1.5	0.2	setosa
4	5.0	3.6	1.4	0.2	setosa
5	5.4	3.9	1.7	0.4	setosa
6	4.6	3.4	1.4	0.3	setosa
7	5.0	3.4	1.5	0.2	setosa
8	4.4	2.9	1.4	0.2	setosa
9	4.9	3.1	1.5	0.1	setosa
10	5.4	3.7	1.5	0.2	setosa
11	4.8	3.4	1.6	0.2	setosa
12	4.8	3.0	1.4	0.1	setosa
13	4.3	3.0	1.1	0.1	setosa
14	5.8	4.0	1.2	0.2	setosa



Dataframes facilitate important data-organizational transformations

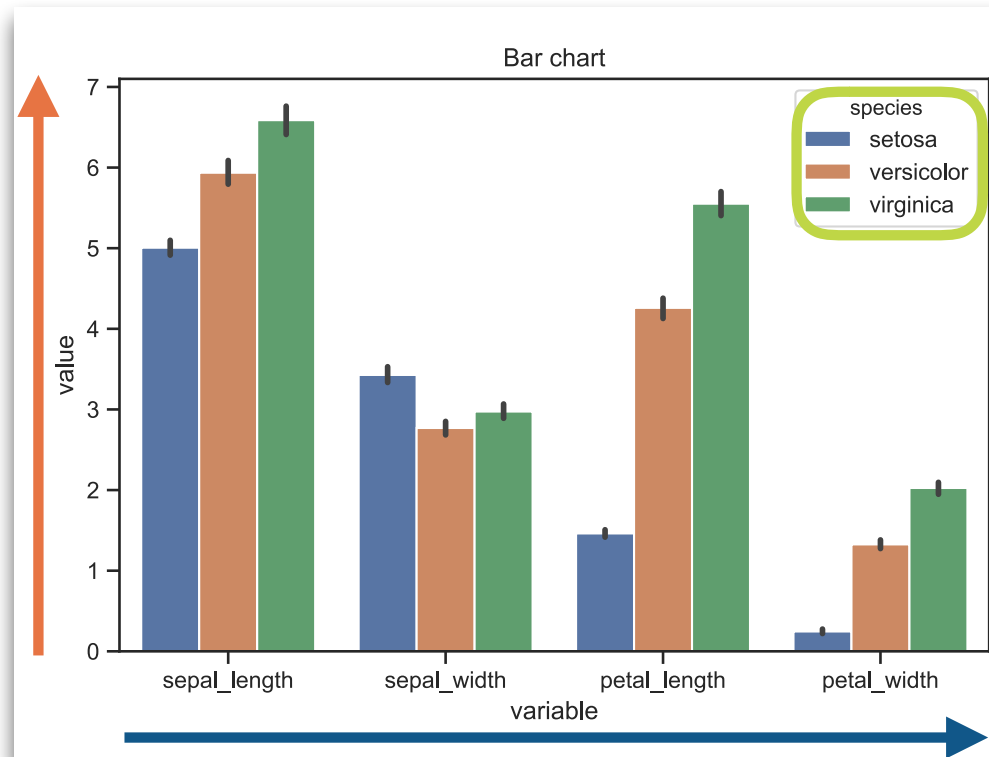
	sepal_length	sepal_width	petal_length	petal_width	species
0	5.1	3.5	1.4	0.2	setosa
1	4.9	3.0	1.4	0.2	setosa
2	4.7	3.2	1.3	0.2	setosa
3	4.6	3.1	1.5	0.2	setosa
4	5.0	3.6	1.4	0.2	setosa

individual = 1 flower

melt

	species	variable	value
0	setosa	sepal_length	5.1
1	setosa	sepal_length	4.9
2	setosa	sepal_length	4.7
3	setosa	sepal_length	4.6
4	setosa	sepal_length	5.0

individual = 1 measurement

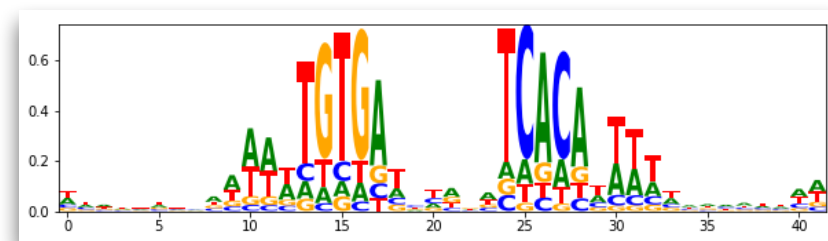


What will we be doing today?

We will start by working with a dataframe listing transcription factor binding sites in the database RegulonDB

	tf	site
0	AcrR	gcgtagattTACATACATTTGTGAATGTATGTAccatagcacg
1	AcrR	cgtgctatggTACATACATTCACAAATGTATGTAaatctaacgc
2	AcrR	catcgggtcaaTTCATTTCATTtgacttatac
3	AcrR	tcactacacgCACATACAACggaggggggc
4	AcrR	atttattaccGTCATTTCATTTCTGAATGTCTGTTtaccctatt
5	AcrR	gctttacctcAAGTTAACTTgaggaattat
6	AcrR	ataattcctcAAGTTAACTTgaggtaaagc
7	Ada	ttcagacgctGCGCTTTGCTTTTCATATTCCGGTTgtcgcgacgg
8	Ada	ggtcaccatcACGCAAAAACCAACAATCTTGCGCttaaatttt
9	Ada	caacaatcttGCGCTTTAATTTTTTTTCGCTGACAaggaagcttt
10	Ada	cgcattacatTGCTGGATAAGAATGTTTTAGCAAatctctttctg
11	AgaR	ttcgtaaaacTTTCGTTTCATTTTCGTTTTGcctattaacg
12	AgaR	ttgcctattaACGCCTTTCTATTAAGCAAatgcaagccca
13	AgaR	tttcagtgcTTTCATTATGTTTCTTTGTgaatcagatc
14	AgaR	aaccattatcTTTCGTTTTATTTTTATCTCaccatgacgc

1. Load TF binding site database a Pandas dataframe
2. Filter for TF of choice
3. Filter for binding sites of the most common length
4. Make a sequence logo



CRP logo (from 358 sites)

We will then parse our computed replication profiles in the form of a data frame

	chromosome	start	stop	reads
0	chrI	1	31	2
1	chrI	32	62	0
2	chrI	63	93	1
3	chrI	94	124	0
4	chrI	125	155	3
5	chrI	156	186	0
6	chrI	187	217	0
7	chrI	218	248	0
8	chrI	249	279	0
9	chrI	280	310	0
10	chrI	311	341	0
11	chrI	342	372	0
12	chrI	373	403	0
13	chrI	404	434	0
14	chrI	435	465	1

1. Load a .bed file as a Pandas dataframe
2. Filter for the chromosome of choice
3. Smooth # reads as a function of position
4. Plot replication profile

