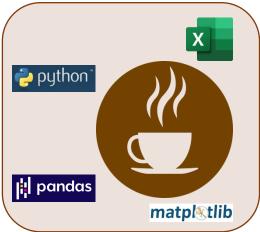
Data Visualization: Excel versus Python When to use which approach ?

University Library Bern, Science Library

Dr. Michael Horn, Coffee & Bit(e)s, Spring 2021



^b UNIVERSITÄT BERN

This Lecture Content

- Excel and Python tools in a nutshell
- data visualization: data size and data complexity
- data visualization: chart/ plot types
- conclusion

Excel and Python tools in a nutshell Excel / Python tools (jupyter, pandas, matplotlib)

➤ Excel:

- Microsoft product
- o desktop application
- o intuitive handling
- multifunctional:
 - spreadsheets for data handling
 - allows complex data analysis
 - allows data visualization
 - .
- has some relevant limitations in comparison to python tools

Python tools:

- o open source: non-proprietary, transparent
- Python:
 - programming (scripting) language
 - easy to learn, very powerful and flexible
- o jupyter notebook:
 - browser-based application
 - integrates code, code output and documentation
- o pandas and matplotlib
 - python libraries (pandas: data processing, matplotlib: data visualization)
 - fast, powerful, flexible, easy to use, very robust

Data size and data complexity Small and simple data sets

^b UNIVERSITÄT BERN

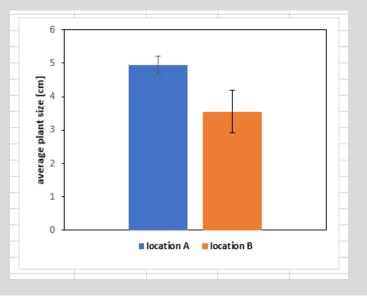
Typical scenario:

X

o manual data acquisitiono rapid data visualization

- (+) very intuitive
- (+) quickly ready for use
- (-) requires intermediate steps (average, stddev)

	A	В	C	
1		plant si	ze [cm]	
2		location A	location B	
3		5.1	3.5	
4		4.9	3	
5		4.7	3.2	
6		4.6	3.1	
7		5	3.6	
8		5.4	4.9	
9				
10				
11	average	4.95	3.55	
12	stddev	0.2629956	0.6396614	
13				



Data size and data complexity Small and simple data sets

^b UNIVERSITÄT BERN

🔁 python' plot = data.mean().plot(kind="bar", color=["blue", "orange"], yerr=[data.std(ddof=0)[0], data.std(ddof=0)[1]], capsize=3) plot.set ylabel("average plant size [cm]", fontsize=14, labelpad=8) import pandas as pd plot.tick params(axis="both", labelsize=14) import matplotlib.pyplot as plt Some pros and cons: plot.tick params(axis="x", labelsize=14, rotation=45) plt.show() data = pd.read excel("simple small ds.xlsx", engine="openpyxl") (+) requires no intermediate data 5 average plant size [cm] steps 4 location A location B (+) faster post processing 0 51 3.5 (axes labels, etc.) 3 1 49 3.0 2 2 4.7 3.2 (-) more intricate 3 4.6 3.1 (-) slower general handling 4 5.0 3.6 0 5 5.4 4.9 location B location A

Data size and data complexity Larger and more complex data sets

^b UNIVERSITÄT BERN

Typical scenario:

manual or automated data acquisitionmore elaborate data visualization

Example data set:

- o Iris data set in csv-format
- o 4 columns with numeric data
- o 1 column with categorical data
- 150 data points (rows)

	-
6.3,3.3,4	7,1.6,Iris-versicolor
7.4,2.8,6	1,1.9,Iris-virginica
6.1,2.6,5	6,1.4,Iris-virginica
4.8,3.4,1	9,0.2,Iris-setosa
5.9,3.2,4	8,1.8,Iris-versicolor
5.4,3.7,1	5,0.2,Iris-setosa
6.0,2.9,4	5,1.5,Iris-versicolor
7.9,3.8,6	4,2.0,Iris-virginica
4.7,3.2,1	3,0.2,Iris-setosa
4.4,3.2,1	3,0.2,Iris-setosa
6.4,3.1,5	5,1.8,Iris-virginica
6.3,2.7,4	9,1.8,Iris-virginica
6.0,2.7,5	1,1.6,Iris-versicolor
6.8,3.2,5	9,2.3,Iris-virginica
5.5,2.3,4	0,1.3,Iris-versicolor

iris dataset shuffled.csv

X					
	A	В	С	D	E
	sepal length	sepal width	petal length	petal width	
1	[cm] 💌	[cm] 💌	[cm] 💌	[cm] 💌	species 💌
2	6.3	3.3	4.7	1.6	Iris-versicolor
3	7.4	2.8	6.1	1.9	Iris-virginica
4	6.1	2.6	5.6	1.4	Iris-virginica
5	4.8	3.4	1.9	0.2	Iris-setosa
6	5.9	3.2	4.8	1.8	Iris-versicolor
			•••		

0 6.3 3.3 4.7 1.6 Iris-version 1 7.4 2.8 6.1 1.9 Iris-virgin 2 6.1 2.6 5.6 1.4 Iris-virgin	🦆 թւ	ython [~]				
1 7.4 2.8 6.1 1.9 Iris-virgin 2 6.1 2.6 5.6 1.4 Iris-virgin		sepal_length (cm)	sepal_width (cm)	petal_length (cm)	petal_width (cm)	species
2 6.1 2.6 5.6 1.4 Iris-virgin	0	6.3	3.3	4.7	1.6	Iris-versicolor
	1	7.4	2.8	6.1	1.9	Iris-virginica
3 4.9 3.4 1.0 0.2 Iric-set	2	6.1	2.6	5.6	1.4	Iris-virginica
3 4.0 5.4 1.5 0.2 115 -560	3	4.8	3.4	1.9	0.2	Iris-setosa
4 5.9 3.2 4.8 1.8 Iris-version	4	5.9	3.2	4.8	1.8	Iris-versicolor

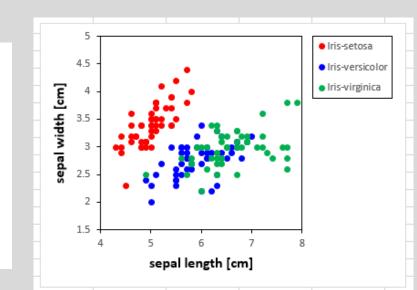
Data size and data complexity Larger and more complex data sets

^b UNIVERSITÄT BERN

Intuitive workflow:

X

- $\circ~$ open csv-file in Excel
- generate three data series corresponding to the categories in the "species"-column (manual filtering, copy-pasting, selecting of data)
- $\circ~$ visualize data



- (+) workflow more intricate, but still intuitive
- (-) requires manual reorganization of the data set
- (-) requires repetitive execution of work steps

Data size and data complexity Larger and more complex data sets

^b UNIVERSITÄT BERN

sepal length (cm)

🔁 python" import pandas as pd import matplotlib.pyplot as plt Some pros and cons: 4.5 df = pd.read csv("iris dataset shuffled.csv", Iris-setosa names=["sepal_length (cm)", Iris-versicolor "sepal width (cm)", Iris-virginica "petal length (cm)", width (cm) 4.0 (+) no repetitive execution of "petal width (cm)", "species"]) manual work steps 3.5 (automated procedure) fig, ax = plt.subplots(figsize=(4,4), nrows=1, ncols=1) colors = {"Iris-setosa":"red", 3.0 "Iris-versicolor":"blue". sepal (+) reusable code: "Iris-virginica":"green"} grouped_dataframe = df.groupby("species") code can be gradually 2.5 for key, data in grouped dataframe: adapted to new sl sw = data.plot(ax=ax, x=0, y=1, kind="scatter", label=key, color=colors[key], s=60) requirements sl_sw.set_xlabel("sepal length (cm)", fontsize=16, 2.0 fontweight="bold", labelpad=8) 5 sl sw.set ylabel("sepal width (cm)", fontsize=16,

fontweight="bold", labelpad=8)

sl sw.tick params(axis="both", labelsize=14)

(1.505, 1.025)).get frame().set edgecolor("black")

sl sw.legend(bbox to anchor=

plt.show()

(-) initial effort is higher

8

Data size and data complexity Very large data sets

b UNIVERSITÄT BERN

Typical scenario:

- o automated data acquisition
- o very large file containing the data set

Example data set:

- o random value data set in csv-format
- o 5 columns with numeric data
- 2`000`000 data points (rows)

very_large_dataset.csv

1.764052345967664,0.4001572083672233,0.9787379841057392,2.240893199201458,1.8675579901499675
 -0.977277879876411,0.9500884175255894,-0.1513572082976979,-0.10321885179355784,0.41059850193837233
 0.144043571160878,1.454273506962975,0.7610377251469934,0.12167501649282841,0.44386323274542566
 0.33367432737426683,1.4940790731576061,-0.20515826376580087,0.31306770165090136,-0.8540957393017248
 2.5529898158340787,0.6536185954403606,0.8644361988595057,-0.7421650204064419,2.2697546239876076
 -1.4543656745987648,0.04575851730144607,-0.1871838500258336,1.5327792143584575,1.469358769900285
 0.1549474256969163,0.37816251960217356,-0.8877857476301128,-1.980796468223927,-0.3479121493261526
 0.15634896910398005,1.2302906807277207,1.2023798487844113,-0.3873268174079523,-0.30230275057533557
 -1.0485529650670926,-1.4200179371789752,-1.7062701906250126,1.9507753952317897,-0.5096521817516535
 -0.4380743016111864,-1.2527953600499262,0.7774903558319101,-1.6138978475579515,-0.2127402802139687
 -0.8954665611936756,0.386902497859262,-0.510805137568873,-1.180632184122412,-0.028182228338654868
 0.42833187053041766,0.06651722238316789,0.3024718977397814,-0.6343220936809636,-0.3627411659871381
 -0.672460447775951,-0.3595531615405413,-0.813146282044454,-1.7262826023316769,0.17742614225375283
 -0.4017809362082619,-1.6301983469660446,0.4627822555257742,-0.9072983643832422,0.05194539579613895
 0.7290905621775369,0.12898291075741067,1.1394006845433007,-1.2348258203536526,0.402341641177549

Data size and data complexity Very large data sets

^b UNIVERSITÄT BERN

X

Intuitive workflow:

- $\circ~$ open csv-file as Excel worksheet
- o visualize data as shown before

- (-) data file cannot be properly opened
- (-) data visualization cannot be performed as expected

	Α	В	С	D	E	F	G	Н	1	J	К	l	
1	ExternalD	ata_1: Getting Data											
2													
3													
4		Microsoft	t Event								V		
5		WICrosoft	LEXCEI								×		
6			The query retu	rned more da	ta than will fit	on a worksho	ot						
7		L 🔶 '	The query returned more data than will fit on a worksheet.										
8			• To continue and display as much of the data as will fit, click OK.										
9			To cancel the query, click Cancel. Change the query in the connection to the data source so that less data is returned.										
10						ОК	Cancel						
11													
12													
10													

Data size and data complexity Very large data sets

^b UNIVERSITÄT BERN

🤁 python"

Some pros and cons:

- (+) data file can be opened without problems
- (+) data visualization can be performed as expected
- (-) procedure can require some computational time (depending on size of data set)

import pandas as pd import matplotlib.pyplot as plt

df_vl =	<pre>df_vl = pd.read_csv('very_large_dataset.csv',</pre>												
df_vl	header=None) df_vl												
	0	1	2	3	4								
	0	1	2	3	4								
0	1.764052	0.400157	0.978738	2.240893	1.867558								
1	-0.977278	0.950088	-0.151357	-0.103219	0.410599								
2	0.144044	1.454274	0.761038	0.121675	0.443863								
3	0.333674	1.494079	-0.205158	0.313068	-0.854096								
4	-2.552990	0.653619	0.864436	-0.742165	2.269755								
1999995	-1.802549	1.129603	-0.480182	0.374793	-0.188933								
1999996	-0.145537	-0.260605	0.886023	1.078775	-0.236004								
1999997	-1.513279	1.294779	0.267679	1.263666	-0.360712								
1999998	-1.401144	1.129229	1.161400	0.484200	1.359349								
1999999	0.590211	0.467616	-1.944715	-0.343985	-0.740190								

2000000 rows × 5 columns

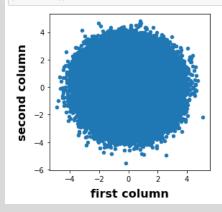


Chart / Plot types

^b UNIVERSITÄT BERN

Choices

X

https://support.microsoft.com/en-us/office/available-chart-types-in-office-a6187218-807e-4103-9e0a-27cdb19afb90#OfficeVersion=Windows

- (+) some relevant chart types are available
- (+) chart recommendation is available
- (-) choices are very restricted
- (-) chart recommendation can be little constructive

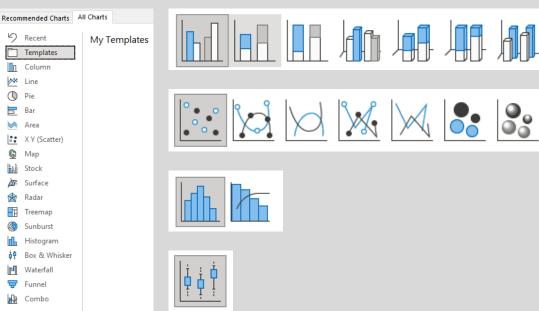


Chart / Plot types Choices



https://matplotlib.org/stable/gallery/index.html

Some pros and cons:

- (+) large amount of chart types available
- (+) good documentation
- (+) good tutorials/ examples
- (-) initial effort is higher (e.g. searching for code)

8	lines, bars and mark	RTR.				1983 I			A.		-	2				0.00
	1111	illii	dhi		After investore of an image	West Barts	Inch	Contras Conno March	Plat 20 state or 20 plat	Detine of 22 fear allowin	Cross 22 Las printe in different planes	Descention philing sectors (instity surveys in 32)	Apple of Station	Anim Styles	Denne (kans Dirid	Asta DeiD
	Revisited Gross	Stated by shall	Grouped has shared with billings	Notice and Date also	1 m	•			-	1	1	-	E3			
			al 🖂	i perhapapata Carpananana	Datus Door	Contrast Unique	Contract Loted Down	Contract Dense	General statutes, plotting stations (level) size res. 31 22 using two represented upper	Propulse and an and an and an and an and an and an	Piloi antines.	Property Beat contribut which a graph	(Ballinin dess	Simony 628 sharesh using 9283am	Dense California of	Column all Annihistor
	Robert Rath	Castlyle	Mailing unique and unitable.	Pailing the selection of test signals		0		-	1	the star	-	Ś				
		\odot		MM		Content and by	Contenting the solution space of splitscoptum.	-	Contain following in a 20 contain plot	II contact	Center 20 Fining an of 20 data	Rearsh in Dava		Derm Sidge Colorine	Green Famil Rice	Chem Inspirate Append
	FBC Dees	Coveralli anar kant	Investeer best veikullen	Louise minarging	-			Change hages with selection		-		•	Controlling for publics and size of colorism with total Ages.			
			*	X		2			Lauren Albiquites	3D ani 3D Aani m sana Pigan	Adecula Ini Othelling	Draw Bel skipski in 30 pist		The second second		
	Land Column Dates	Everytel Down	* Bud polyger	Pill Bolance and April	Bridge Direse	Contraction of	Logo Secondara	Mined Intersportency with under to 20 Integer	-		36				Slider Roam For Holes United Antegril	
1				mhar		333333			Germaliz polypors is 18 unite 22 los: graph	22 option plat	Relating in 22 pilot	32 sustinger	1-++	•	in the second	
	Piling for one for the second		Paint Hart	Revisitati	Multipleg for	Angeleters for Television	Contrast patient Singularity spatient aller	Lager Images	Ð	s i 🗢	:چ	۲	Anasta Dagir?	Baather Histogram (Lanstable Barn)	Respirations Autom	Emple fam Deiler
	1111-								10 sime	22 plans on samplets.	20 satisse (side risp)	All statutes (solid) address			• •	
		Constitution discrimination on historical law shart		Decision of the second	Mainlaw	Multi berger	Pade Gree	patienend gibt	٠	•	働	1	Emple Ages Division 3	Bingin ImageCod	Strade Integration 2	Toph Asidovi
		Tanges			22		::::		25 without (Analyticated)	Stratini alli pila incelluito.	Tend over the factores in TED	Trangalar, 22 metano pira				
		1-7188	EN THE		-	Emerylid	Dualding Same	Advanced policy and spacetry bindlers	*		ê) 🔶	~	Bright Galaxie			
			Valuet and Law					an an	Triangulas TG (Brd annihier yiki	Tingale 32 tarlans	More the spalle 32	32 mail when the	Axis Artist	ranna		1-1-1-1-1-1-1-1-1-1-1-1-1-1-1-1-1-1-1-
		103000		17.75						۲	4		Ant Consister	SEE.	Ann line spin	Sector pitter
	Reality Context	Rate Grad	Anter de est	14532 Anton Vanna		Transa Areas Determiny		Department from	3D mail phi si the many lege	3D meet industries and will up online	20 scienti / columnicia plat adli oglivation insecilention	32) environe plut				28 23+
				102044					26	10.10				000		
	Sautor side with pie-	COELCI Marter receptor	Trades Typest		A A	Episado Como	Table Dates	$\Delta \Lambda$	Proving 32 minimum plat	12 workers plate is not detailed			Down Devel Jose Grag	mill, herb D. annuellel. Braham,	a Generating Ass	Paraticipation
	3/17/1		.		Subplots, some and t	Igares			-17		hth		\mathbb{Z}		UDD	
			-	Restances Analysis		Annihis agent	www.		9 <u>/</u>	Community of the second	Log Ter	LagGene	Paraste sale deces	Tableid algerowi	Publical direction	Emple fan Denslandt
				14774				Controlling using British accept reading and ableg object								4
	And and			Crating a bandra with time, dates, and inst		Ann dan Khai			Lagram .	Logi Gross	Lapinory manufactures	100	Bengle Auto Dissuitor(D	Emple Ant. Part	Reeple frameted?	E-ph-Andre
	ALL SCHOOL			with lines, dates, and lines	Jam Pegn	fam. Dans 27ml	adiapat Gross	Reput anti-mprod athi			an and to have					

^b UNIVERSITÄT BERN

Chart / Plot types Usability

Use case:

- histogram of the Iris sepal length data
- $\circ~$ separated by species
- $\circ~$ integrated in one single chart

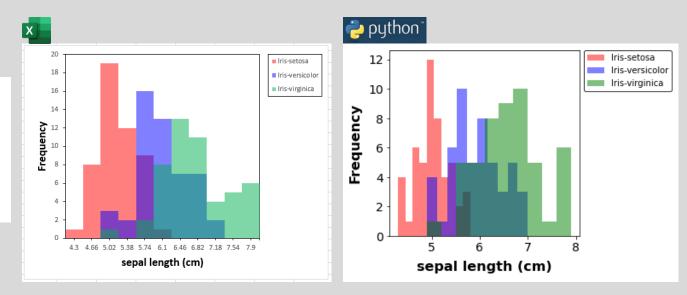


Chart / Plot types Usability



Workflow in Excel:

- o open csv-file in Excel
- generate three data subsets corresponding to the categories in the "species"-column (manual filtering, copy-pasting, selecting of data)
- generate a set of appropriate bins (manual)

- load and activate the Analysis ToolPak in Excel
- from the ToolPak use the analysis tool *Histogram* to calculate the bin frequencies for each data subset

 visualize the frequencies as individual data sets in a bar plot Some problems with this workflow:

- o very laborious and inefficient
- does not actually generate a histogram, but a bar plot

X

See: https://superuser.com/questions/1310668/how-do-i-overlay-two-histograms-in-excel

Chart / Plot types Usability



🦆 python"

Workflow in Python: minor changes to the existing code for the scatter plot \rightarrow code for the histogram.

Code for scatter plot

```
fig, ax = plt.subplots(figsize=(4,4), nrows=1,
                       ncols=1)
colors = {"Iris-setosa":"red",
          "Iris-versicolor":"blue",
          "Iris-virginica":"green"}
grouped dataframe = df.groupby("species")
for key, data in grouped dataframe:
    sl_sw = data.plot(ax=ax, x=0, y=1, kind="scatter",
                      label=key, color=colors[key], s=60)
   sl sw.set xlabel("sepal length (cm)", fontsize=16,
                     fontweight="bold", labelpad=8)
    sl sw.set ylabel("sepal width (cm)", fontsize=16,
                     fontweight="bold", labelpad=8)
    sl sw.tick params(axis="both", labelsize=14)
    sl sw.legend(bbox to anchor=
    (1.505, 1.025)).get frame().set edgecolor("black")
plt.show()
```

Code for histogram

```
fig, ax = plt.subplots(figsize=(4,4), nrows=1,
                       ncols=1)
colors = {"Iris-setosa":"red",
          "Iris-versicolor":"blue",
          "Iris-virginica":"green"}
grouped dataframe = df.groupby("species")
for key, data in grouped dataframe:
   sl 10 = data["sepal length (cm)"].plot(ax=ax, kind="hist",
                      label=key, color=colors[key],
                      alpha=0.5, bins=10)
    sl 10.set xlabel("sepal length (cm)", fontsize=16,
                     fontweight="bold", labelpad=8)
    sl 10.set ylabel("Frequency", fontsize=16,
                     fontweight="bold", labelpad=8)
    sl 10.tick params(axis="both", labelsize=14)
    sl 10.legend(bbox to anchor=
    (1.505, 1.025)).get frame().set edgecolor("black")
plt.show()
```



Conclusion: DV Excel versus Python tools When to use which approach ?

> Excel:

- o small and simple data sets
- o basic chart types
- o little complex use cases
- o for preliminary estimations
- o for informal presentations/ meetings
- if use of code should be avoided

Python tools:

- o larger and more complex data sets
- o more intricate chart types
- complex use cases with more elaborate data visualization
- for data projects reaching a more mature level
- o for more formal presentations/ meetings
- o for publications

Thank you for your attention

b UNIVERSITÄT BERN

University Library Bern, Science Library

Dr. Michael Horn, Coffee & Bit(e)s, Spring 2021



→ see «Coffee & Bit(e)s» for lecture notes

<u>https://github.com/ubnpl/Coffee_and_Bites/tree/main/2021_DV_Excel_vs_Python</u>
 see «Notebook_DV_Excel_vs_Python_2021.ipynb» for the Python code used

