

Digital humanities

Cvičení: Práce s digitálními daty (Jupyter Lab)

Jindřich Marek

Přehled

- toto cvičení seznamuje s programovacím jazykem Python a Jupyter notebooky
- pracujeme s historickými daty o pražské univerzitě (1433-1622)

Cíle cvičení

- po dokončení tohoto cvičení budete schopni
 - nastavit a spustit Jupyter notebook
 - psát a spouštět kód pro Python v notebooku
 - načítat data a manipulovat s nimi pomocí pandas (modul pro Python)
 - vytvářet interaktivní tabulky a vizualizace
 - pochopit základní strukturu datové analýzy v Pythonu

Požadované nástroje

- JupyterLab na univerzitním serveru
 - <https://hpc.troja.mff.cuni.cz:8000>
 - přihlaste se prostřednictvím CAS



Dataset: pražská univerzita (1433-1622)

- zdroj: článek Františka Šmahela a Miroslava Truce
- obsah: historická data o promociích na pražské univerzitě a o počtu učitelů
- formát: XLSX (Excel)
- klíčové sloupce
 - rok
 - počet bc. promocí
 - počet mgr. promocí
 - počet učitelů

Postup

- instalace knihoven - nastavení prostředí
- načtení dat - import z XLSX souboru
- přehled o datech - základní statistiky
- transformace - převod tabulky ze širokého na úzký formát
- interaktivní vizualizace - grafy v Plotly Express

Zdroj dat

Šmahel, František – Truc, Miroslav:
Studie k dějinám University Karlovy v letech 1433–1622.

In: Acta Universitatis Carolinae: Historia Universitatis Carolinae Pragensis:
Příspěvky k dějinám Univerzity Karlovy, roč. 4, č. 2, 1963, s. 3–60.

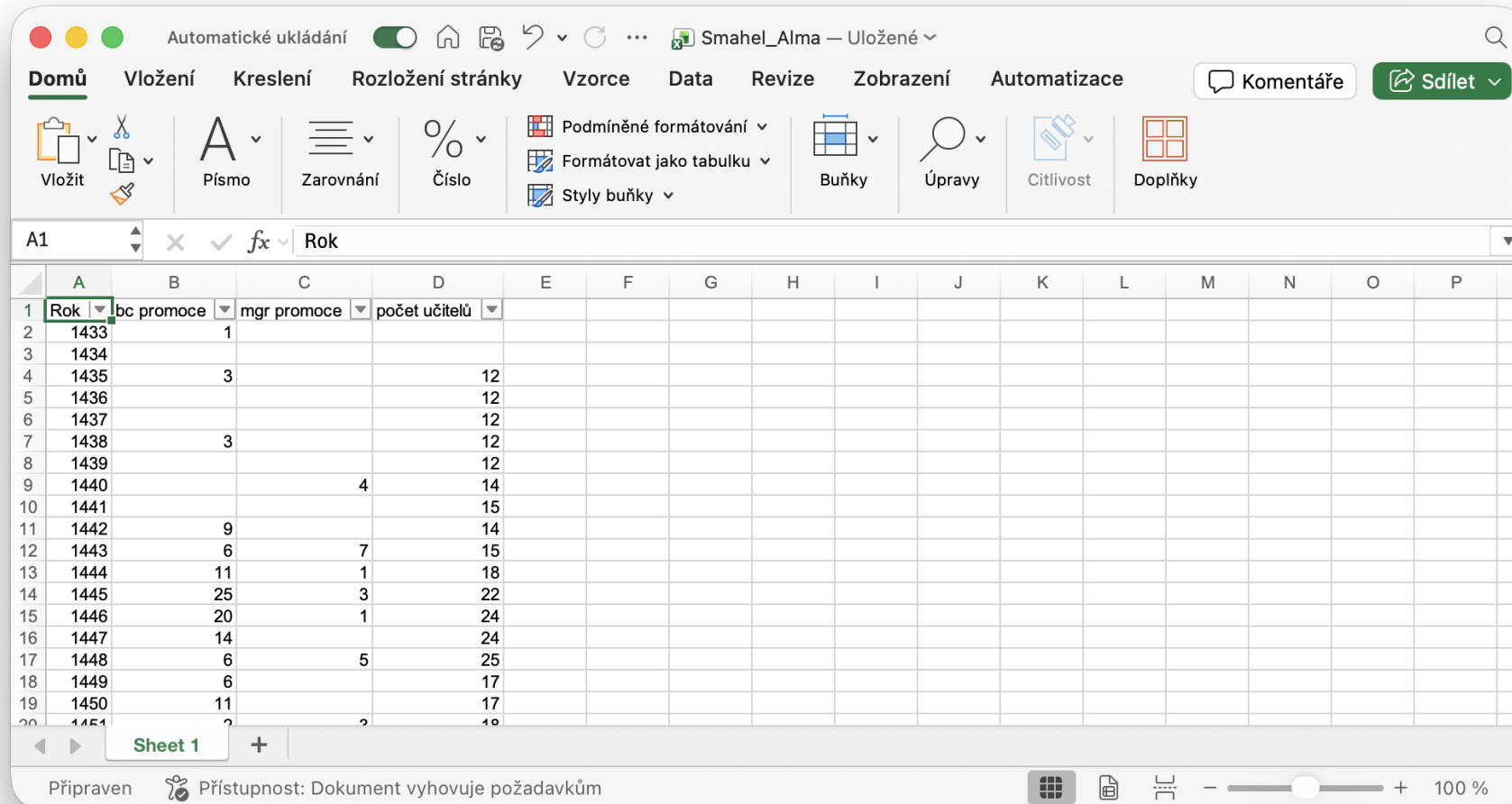
Počet učitelů, bakalářských a mistrovských promócí na Univerzitě Karlově v letech 1433–1622

Rok	BP	MP	U	Rok	BP	MP	U	Rok	BP	MP	U	Rok	BP	MP	U
1433	1	-	?	1481	11	4	17	1529	-		7	1577	28	-	12
1434	-	-	?	1482	-	-	16	1530	13	4	7	1578	20	-	9
1435	3	-	12	1483	-	-	17	1531	-	-	7	1579	-	8	8
1436	-	-	12	1484	6	-	17	1532	13	-	9	1580	41		11
1437	-	-	12	1485	-	4	16	1533	-	-	6	1581	20	-	8
1438	3	-	12	1486	8	-	17	1534	29	5	7	1582	-	8	6
1439	-	-	12	1487	14	-	16	1535	12	-	7	1583	28	-	8
1440	-	4	14	1488	6	-	16	1536	-	-	9	1584	-	11	6
1441	-	-	15	1489	16	-	15	1537	17	-	9	1585	31	-	10
1442	9	-	14	1490		-	14	1538	-	-	9	1586	-	-	?
1443	6	7	15	1491	9	-	14	1539	-	-	9	1587	-	-	?
1444	11	1	18	1492	11	-	14	1540	26	-	8	1588	-	10	?
1445	25	3	22	1493	11	6	17	1541	-	9	8	1589	37	-	?
1446	20	1	24	1494	-	-	15	1542	-	-	9	1590	-	-	?
1447	14	-	24	1495	15	-	17	1543	30	-	9	1591	38	-	7

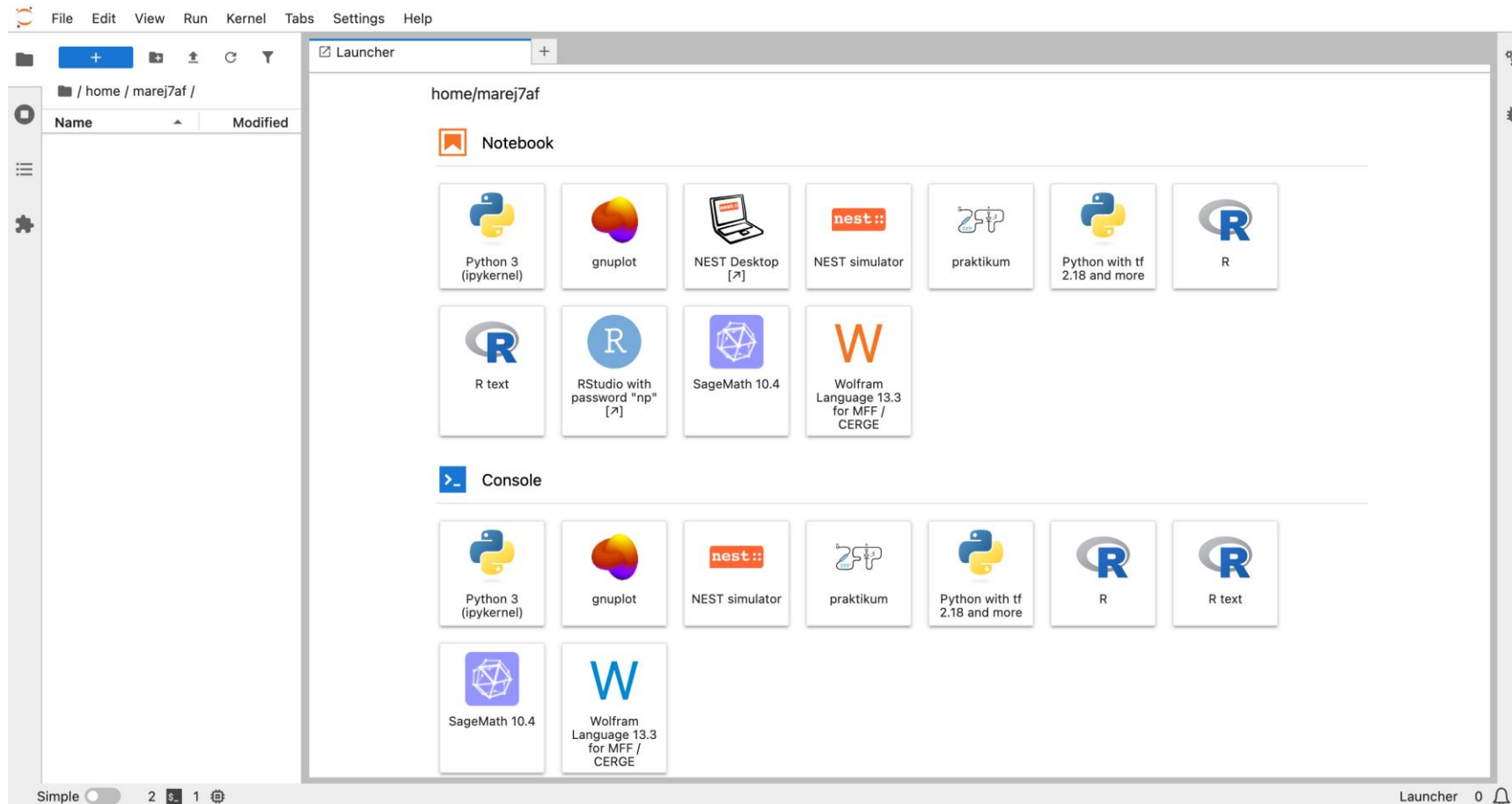
1448	6	5	25	1496	-	-	15	1544	17	-	9	1592	-	-	?
1449	6	-	17	1497	10	-	14	1545	-	9	8	1593	47	10	?
1450	11	-	17	1498	7	-	13	1546	28	-	10	1594	-	-	?
1451	2	3	18	1499	12	2	13	1547	-	-	9	1595	52	-	?
1452	8	-	17	1500	17	-	13	1548	24	-	9	1596	-	12	?
1453	6	-	20	1501	-	-	13	1549	-	-	9	1597	46	-	?
1454	17	6	18	1502	19	3	12	1550	24	4	7	1598	-	-	?
1455	2	-	15	1503	-	-	12	1551	-	-	7	1599	32	-	9
1456	7	4	21	1504	19	-	12	1552	11	10	7	1600	-	8	10
1457	15	-	18	1505	-	3	12	1553	20	-	9	1601	46	-	10
1458	14	11	19	1506	19	-	13	1554	-	-	8	1602	-	-	9
1459	6	-	22	1507	-	-	12	1555	33	-	9	1603	46	-	10
1460	16	4	21	1508	6	-	13	1556	-	2	8	1604	-	13	8
1461	10	-	21	1509	10	-	12	1557	-	-	8	1605	62	-	8
1462	13	9	23	1510	-	-	11	1558	38	-	7	1606	-	11	9
1463	11	-	22	1511	17	1	12	1559	-	-	8	1607	57	-	8
1464	-	-	21	1512	4	-	11	1560	24	-	6	1608	-	12	8
1465	10	4	17	1513	-	5	11	1561	-	8	7	1609	55	-	8
1466	8	-	15	1514	-	-	11	1562	25	-	8	1610	-	12	10
1467	14	7	16	1515	35	4	11	1563	-	-	7	1611	61	-	8
1468	13		15	1516	16	-	11	1564	19	-	8	1612	38	11	10
1469	-	-	13	1517	13	-	12	1565	-	7	7	1613	17	8	?
1470	5	6	18	1518	-	7	12	1566	28	-	6	1614	33	13	?
1471	-	-	17	1519	15	-	11	1567	-	-	9	1615	32	13	?
1472	9	-	16	1520	5	-	12	1568	30	-	9	1616	29	15	?
1473	-	-	14	1521	17	-	10	1569	-	6	8	1617	27	9	?
1474	11	-	14	1522	10	5	11	1570	26	-	8	1618	39	-	?
1475	-	4	14	1523	19	-	10	1571	15	-	7	1619	52	19	?
1476	14	-	15	1524	8	6	8	1572	-	7	7	1620	57	-	?
1477	-	-	15	1525	-	-	9	1573	21	-	8	1621	-	-	?
1478	6	-	16	1526	13	-	8	1574	-	-	8	1622	-	-	?
1479	12	-	16	1527	-	-	9	1575	37	-	8				
1480	10	-	18	1528	24	-	9	1576	-	8	7				

BP – bakalářské promoce, MP – mistrovské promoce, U – učitelé

Data ve formátu MS Excel



JupyterLab



Návod k prostředí JupyterLab

- The JupyterLab Interface
 - <https://jupyterlab-exp.readthedocs.io/en/latest/user/interface.html>
- Working with Files
 - <https://jupyterlab-exp.readthedocs.io/en/latest/user/files.html>

Nahrání notebooku na server

- stáhněte si následující soubory z Moodle
 - Smahel_Alma.ipynb (notebook)
 - Smahel_Alma.xlsx (zdrojová data)
- přetáhněte je v JupyterLabu do okna vlevo (nadpisy Name, Modified)

File

Edit

View

Run

Kernel

Tabs

Settings

Help

+

/ home / marej7af /

Name

Modified

Smahel_Alma.ipynb

now

Smahel_Alma.xlsx

2s ago

Launcher

home/marej7af

Notebook

Python 3
(ipykernel)

gnuplot

NEST Desktop
[?]

NEST simulator

praktikum

Python with tf
2.18 and more

R

R text

RStudio with
password "np"
[?]

SageMath 10.4

Wolfram
Language 13.3
for MFF /
CERGE

Console

Python 3
(ipykernel)

gnuplot

NEST simulator

praktikum

Python with tf
2.18 and more

R

R text

SageMath 10.4

Wolfram
Language 13.3
for MFF /
CERGE

Simple

2

1

Launcher

0

Otevření notebooku

- poklepáním otevřete soubor
 - Smahel_Alma.ipynb



/ home / marej7af /

Name	Modified
• Smahel_Alma.ipynb	now
Smahel_Alma.xlsx	1s ago



Launcher Smahel_Alma.ipynb +

+ - Copy Paste Run Cell Run All Markdown ↺

Notebook Python 3 (ipykernel) ○

▼ Interaktivní vizualizace dat v Pythonu

Instalace potřebných knihoven

Každou buňku s kódem je možné spustit, když kliknete do ní a poté nahoře na tlačítko ►, případně je možné použít tlačítko ►► (to navíc restartuje kernel; poté spustí buňky jednu po druhé).

Buňky v notebooku byste měli spouštět v pořadí, jak jsou v notebooku, protože buňky níže často pracují s výsledkem z buněk výše.

Vedle buněk s kódem jsou zde buňky s markdownem (jako je tato); ty je možné editovat po poklepání (ale přirozeně v nich nelze nic spouštět).

Pokud nefunguje spuštění dalšího kódu po spuštění a provedení instalace v první buňce, restartujte jádro z menu (Kernel / Restart kernel...).

```
[6]: %pip install -Uqq pandas openpyxl itables plotly
```

Note: you may need to restart the kernel to use updated packages.

Načtení dat

Tato buňka načte data z excelového souboru, který jsme přiložili.

Poté zobrazí několik prvních řádků.

Na začátku knihovny je import knihovny `pandas`. Importujeme ji pod krátkým názvem `pd` a pak odkazujeme na její funkce, např. `pd.read_excel`.

Data máme načtená v proměnné `data`. Jedná se o datový typ `pandas DataFrame`.

```
[7]: import pandas as pd

# Načtení dat
file_path = "Smahel_Alma.xlsx"
```

Spuštění buněk v notebooku

- sledujte instrukce přímo v notebooku
- klíčové je spuštění buňky s instalací balíčků:
 - `pip install jupyter pandas openpyxl itables plotly`
- výsledný graf by měl vypadat jako ten na následujícím snímku

Množství bakalářských a magisterských promočí a počet učitelů na Univerzitě Karlově v letech 1433–1622

