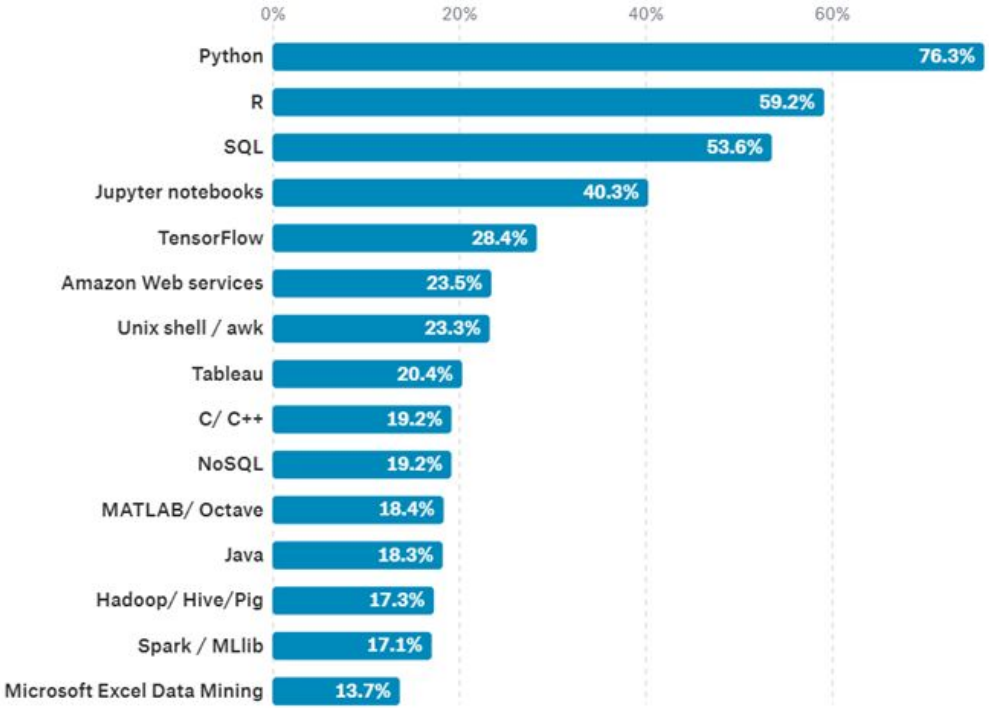
Nástroje na dátovú analýzu

R vs. Python

VOS2, 26.9.2018

R a Python sú aktuálne najpoužívanejšie jazyky na dátovú analýzu



Zdroj: The State of Data Science & Machine Learning. 2017. Kaggle. https://www.kaggle.com/surveys/2017

Formát dátovej analýzy = notebook

Jupyter Notebook, Jupyter Lab

- Klient-server
- Zobrazovaný a interpretovaný v prehliadači
- Interne reprezentovaný ako JSON (nevýhoda z pohľadu verziovania)
- Colab

RMarkdown

- Markdown s pridanými blokmi R kódu
- Kompilovanie do iných formátov (HTML) prostredníctvom knižnice *knitr*
- Simulovanie notebook funkcionality v prostredí RStudia

IDE

PyCharm

- https://www.jetbrains.com/pycharm/

Spyder

- https://www.spyder-ide.org/

RStudio

- https://www.rstudio.com/

Reprezentácia údajov = Data Frame

Pandas

pandas.Series

pandas.DataFrame

data.frame

```
df <- data.frame(
     name = c("Robo", "Jakub"),
     has_phd = c(TRUE, TRUE),
     sex = factor(c("male", "male")))
```

Filozofia jazyka

- General purpose OO jazyk
- Multiparadimový jazyk (základ v C++)
  - Funkcionálne prvky
- Pythonic way (Zen of Python)
  - https://www.python.org/dev/peps/pep-0020/

- Vznikol ako dialekt štatistického jazyka S
- Základná dátová štruktúra je *vektor*
- Multiparadigmový jazyk
  - Procedurálny základ
  - Objektovo-orientované črty (generické funkcie, typy)
  - Funkcionálne črty - funkcia ako first-class citizen

Užitočné knižnice

- Numpy, scipy
- Pandas
- Matplotlib, seaborn
- Scikit-learn
- Nltk
- Gensim
- ...

- Plyr
- Dplyr
- Tidyr
- Lubridate
- Data.table
- Ggplot2
- Caret
- ...

Výkonnosť manipulácie s dátami: data.table vs. dplyr vs. pandas

Zdroj:
https://github.com/Rdatatable/data.table/wiki/Benchmarks-:-Grouping

https://h2oai.github.io/db-benchmark/

Zhrnutie

Plusy

- Konzistentnosť knižníc, STD knižnica
- Dokumentácia
- Testovanie + debugovanie
- Čitateľnosť
- Normálny general-purpose OOP jazyk

Mínusy

- Python 2 alebo Python 3?
- Stále chýba veľa R modelov
- Znalosť programovania výhodou

Plusy

- Veľa existujúcich modelov a knižníc
- Komunita
- Dokumentácia
- ggplot2
- dplyr, data.table

Mínusy

- Vyvinuté štatistikmi
- Čitateľnosť / konzistentnosť kódu

Užitočné zdroje

- http://pandas.pydata.org/pandas-docs/stable/10min.html
- Exploratory computing with python http://mbakker7.github.io/exploratory_computing_with_python/
- Galéria Jupyter notebookov / kníh https://github.com/ipython/ipython/wiki/A-gallery-of-interesting-IPython-Notebooks
- Intro do štatistiky https://github.com/rouseguy/intro2stats
- Titanic na Kaggle https://github.com/agconti/kaggle-titanic/blob/master/Titanic.ipynb

- Coursera Data Science Specialization https://www.coursera.org/specialization/jhudatascience/1
- https://www.youtube.com/user/rdpeng
- Advanced R http://adv-r.had.co.nz/
- R for Data Science http://r4ds.had.co.nz/index.html
- http://swirlstats.com/
  - Kurz “R Programming E”: lekcie 1-9 + 12
- data.table https://cran.r-project.org/web/packages/data.table/vignettes/datatable-intro.html

Zadanie do budúceho týždňa (odovzdanie do AIS)

Spraviť cvičenia v Pythone **[10b]**

https://github.com/robom/vos2-2018/blob/master/python-exercises.zip

Numpy **[optional]**
https://github.com/rougier/numpy-100

Pandas **[optional]**
https://github.com/ajcr/100-pandas-puzzles

Spraviť R tutoriál **[10b]**
https://www.datacamp.com/courses/free-introduction-to-r