LO5

Prepare Data for Modeling

Topics cover in LO5 & LO6

- Introduce various python libraries
- Filtering and selecting data
- Concatenating and transforming data
- Data visualization best practices
- Visualizing data
- Creating a plot
- Creating statistical data graphics
- Performing basic math and linear algebra
- Correlation analysis
- Multivariate analysis
- Data sourcing via web scraping

Objective of today's session

After attending this session, you should know

- Panda library introduction
- Filtering and selecting
- Treating missing values

Coding languages for data science

- Python
- R
- Julia
- Go

- Python is a high-level interpreted coding language that's useful for a wide variety of applications.
- It is an official programming language of Google

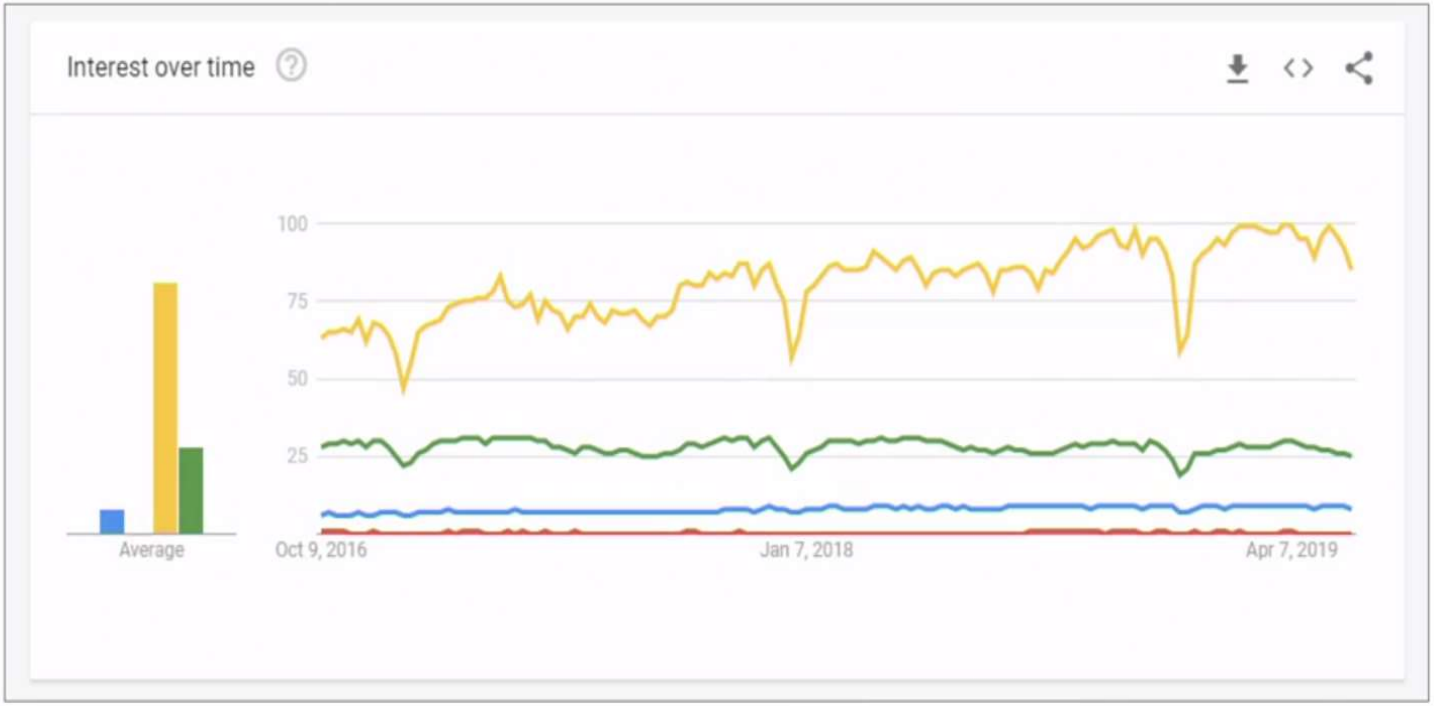
Benefits of using Python

- It is extremely easy to learn and it's human readable.
- Got an extensive array of well-supported date science libraries.
- Got the biggest user base of all data science languages.
- Use for building predictive web applications as well and use for lot of different functions, not just data science.

Python is a popular language

Google Trends:

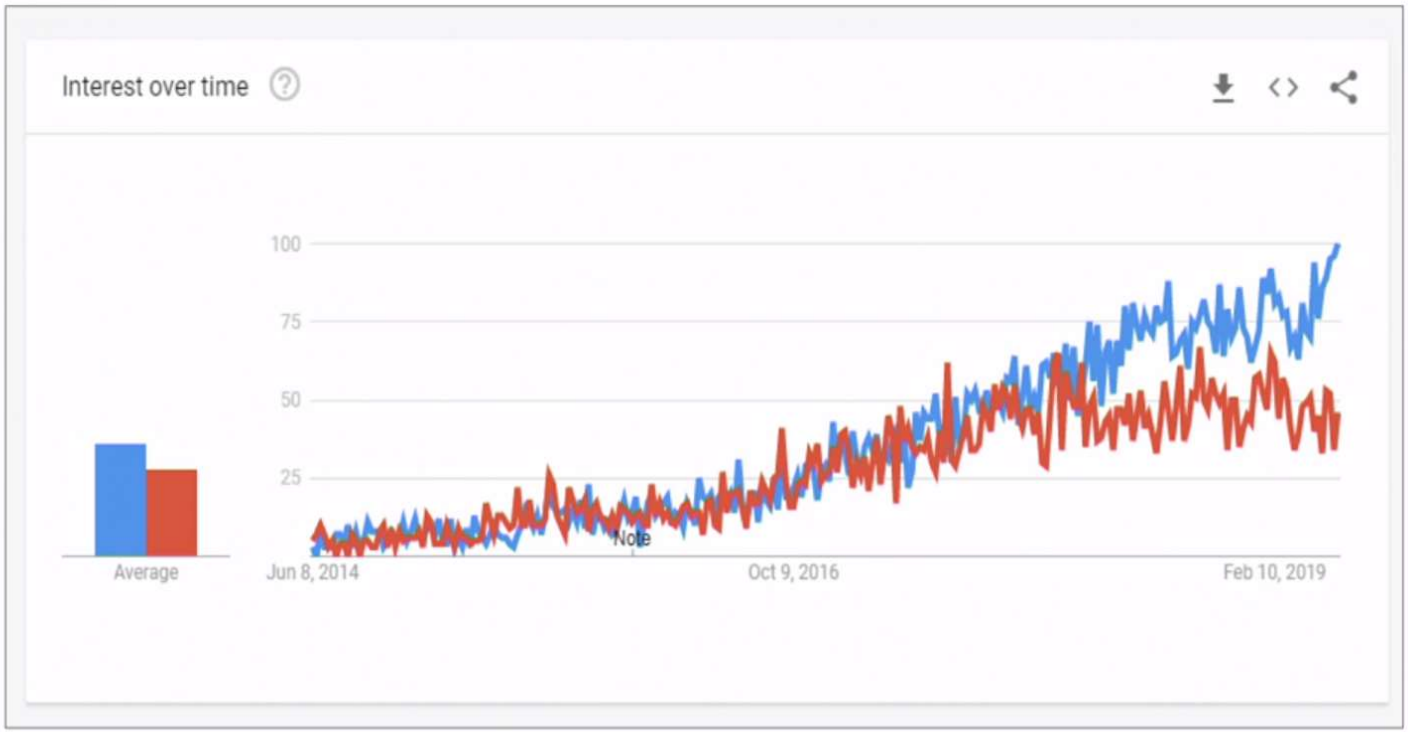
1. "Python"
2. "R"
3. "Go"
4. "Julia"



Python is most popular in data science

Google Trends:

1. "Python for data science"
2. "R for data science"



Why use Python for working with Data

Python is useful for:

- Data science, data analytics, and data engineering
- Useful in both a professional and an academic environment
- Python is an open-source programming language
- Web development
- Application development
- Game development

Main Python libraries for data science

Advanced Data Analysis	Data Visualization	Machine Learning
NumPy SciPy pandas	Matplotlib Seaborn	scikit-learn TensorFlow Keras

Pandas library introduction

- Pandas is useful for its fast data cleansing preparation, powerful analysis capabilities, ease of use for data visualization, ease of use for machine learning
- its compatibility with NumPy array and matrices.
- It is built on top of NumPy.
- Arrays and matrices are called series and DataFrames in pandas.

Shortcuts in jupyter: https://yoursdata.net/jupyter-lab-shortcut-and-magic-functions-tips/

Indexing in pandas

- An index is a list of integers or labels you use to uniquely identify rows and columns.

We use

- A set of square-brackets [.....]
- The .loc[] indexer

Introducing the pandas library

- A DataFrame object is pretty much a spreadsheet of rows and columns
- the rows and columns individually are actually series objects in the pandas library
- DataFrames are indexable.
- A series object is a single row or column and it is always indexed

	column 1	column 2	column 3	column 4	column 5	column 6
row 1	0.228273	1.026890	-0.839585	-0.591182	-0.956888	-0.222326
row 2	-0.619915	1.837905	-2.053231	0.868583	-0.920734	-0.232312
row 3	2.152957	-1.334661	0.076380	-1.246089	1.202272	-1.049942
row 4	1.056610	-0.419678	2.294842	-2.594487	2.822756	0.680889
row 5	-1.577693	-1.976254	0.533340	-0.290870	-0.513520	1.982626
row 6	0.226001	-1.839905	1.607671	0.388292	0.399732	0.405477

Comparison operators in pandas

Operator	Arithmetic Operation
==	True if values of two operands are equal
!=	True if values of two operands are unequal
<>	True if values of two operands are unequal
>	True if the left operand has a value that's greater than the right operand
<	True if the left operand has a value that's less than the right operand
>=	True if the left operand has a value that's greater than or equal to the right
<=	True if the left operand has a value that's less than or equal to the right operand

Code demonstration

- Introduce Jupyter notebook
- Plain indexing
- Data slicing
- Arithmetic comparisons

PACKAGES/MODULE: https://ajaytech.co/2020/04/21/modules-vs-packages-vs-libraries-vs-frameworks/

Random seed: https://www.youtube.com/watch?v=8B1z3xwNy2s

Data types: https://jakevdp.github.io/PythonDataScienceHandbook/02.01-understanding-data-types.html

Summary

- Panda library introduction
- Filtering and selecting
- Treating missing values

Himanshu Patel, Instructor
Saskatchewan Polytechnic
email: patelh@saskpolytech.ca
Mining building, Saskatoon