多元统计分析—R 与 Python 的实现

引言

吴喜之

June 28, 2019

多元分析的对象

- 数据形式: 规范的横截面数据, 观测值及变量可以按照行列排成方阵.
- 变量: 变量都是相关的, 而且每个观测值都是同时度量的.

根据机器学习术语,

- 有指导 (或有监督) 学习 (supervised learning): 是有因变量作为目标变量, 主要目的是建立可以预测的模型, 具体要介绍回归和分类
- 无指导 (或无监督) 学习 (unsupervised learning): 没有目标变量, 目的是揭示相关变量所提供的潜在重叠信息所代表数据的背景结构, 因此, 简化、降维、汇总是其主要特点.

根据因变量的性质, 有监督学习分为两种:

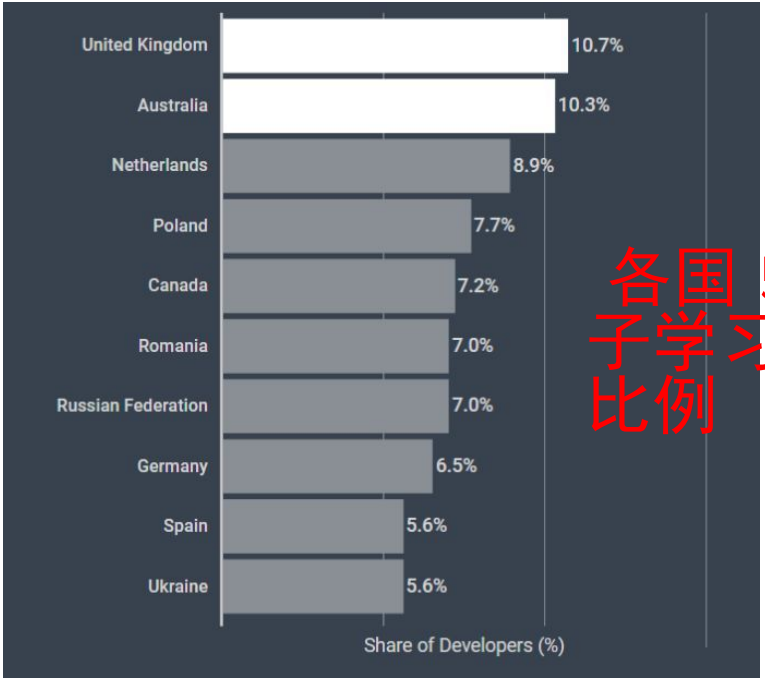
- 因变量为数量变量时有监督学习称为回归 (regression). 本书将要介绍的回归方法包括决策树及以决策树为基础的组合学习方法, 包括随机森林、bagging 等; 此外还介绍经典的线性回归等.
- 因变量为分类变量时有监督学习称为分类 (classification). 本书将要介绍的分类方法包括决策树及以决策树为基础的组合学习方法 (ensemble learning method), 包括 bagging、随机森林、adaboost; 还有二分类的 logistic 回归.

无监督学习 (传统的多元分析内容), 按照模型为:

- 主成分分析
- 因子分析
- 聚类分析
- 典型相关分析
- 对应分析
- 多维尺度分析

不会编程将等同于文盲

2019 年 3 月 13 日，教育部公布了《2019 年教育信息化和网络安全工作要点》，明确表示，今年将启动中小学生信息素养测评，并推动在中小学阶段设置人工智能相关课程，逐步推广编程教育，还将编制《中国智能教育发展方案》。



各国 5-10 岁孩子学习编程的比例

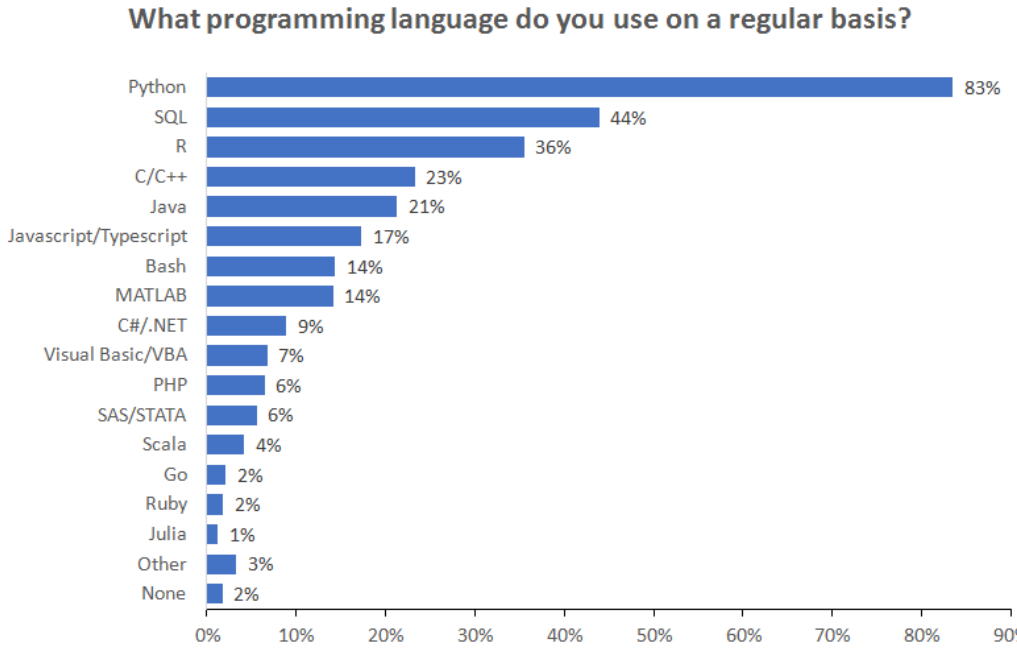
学什么编程语言？R 还是 Python?

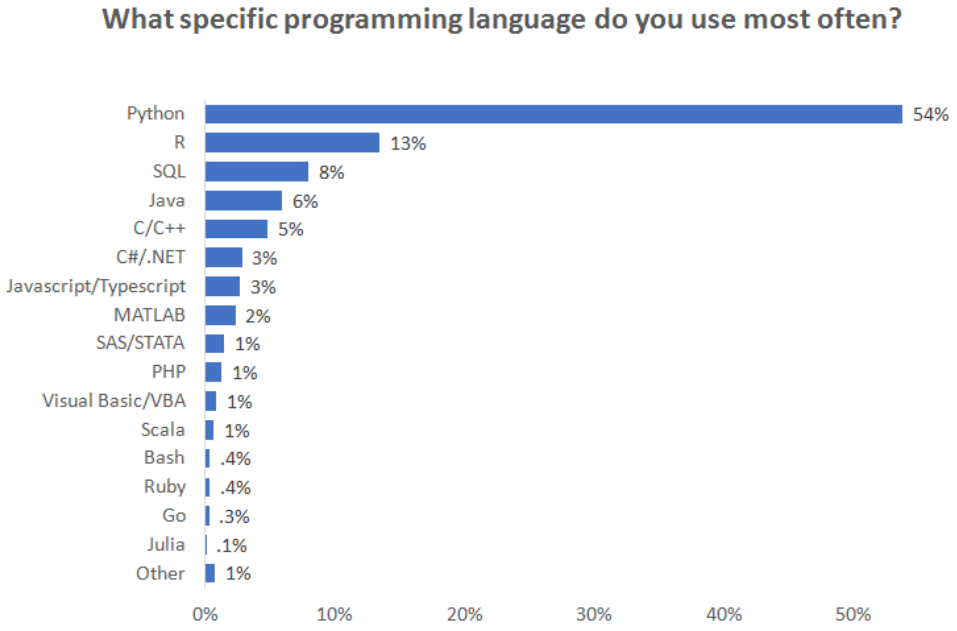
- 学习数据科学，而不仅仅是编程本身
- **R** 和 **Python** 功能不同
- 不要选择，**R** 和 **Python** 两个都学
- 能够加强你的科学对话能力及提升你的数据科学生涯
- 学编程语言并不难，通过分析数据去学，而不是通过语法说明

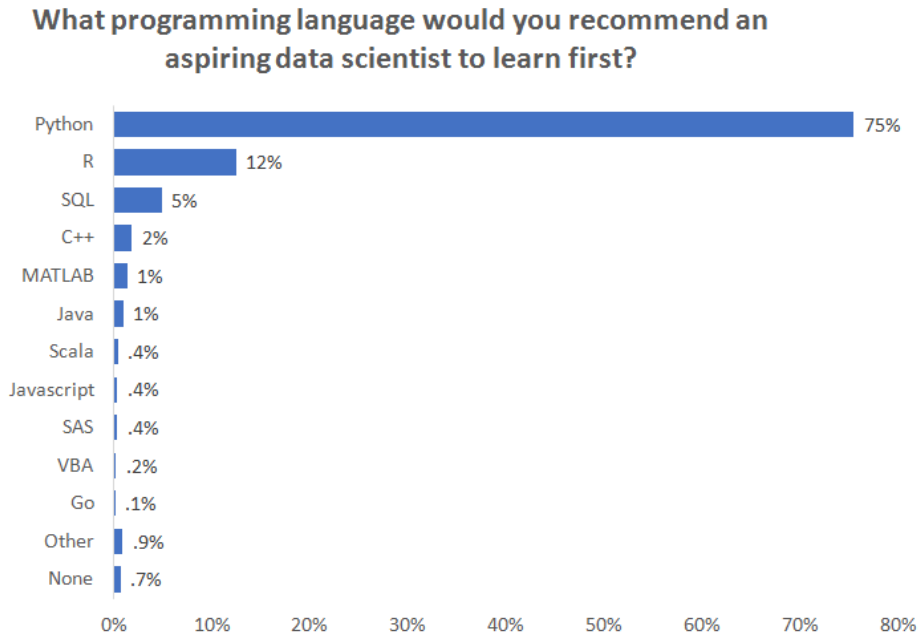
Kaggle 最近对近 24000 名数据专业人员进行的调查显示[1], Python, SQL 和 R 是最流行的编程语言. 到目前为止, 最流行的是 Python（83%使用率）。四分之三的专业人士建议有抱负的数据科学家首先学习 Python. [2]

[1] https://www.kaggle.com/kaggle/kaggle-survey-2018

[2] http://customerthink.com/programming-languages-most-used-and-recommended-by-data-scientist/


What programming language do you use on a regular basis?
Python 83%
SQL 44%
R 36%
C/C++ 23%
Java 21%
Javascript/Typescript 17%
Bash 14%
MATLAB 14%
C#/.NET 9%
Visual Basic/VBA 7%
PHP 6%
SAS/STATA 6%
Scala 4%
Go 2%
Ruby 2%
Julia 1%
Other 3%
None 2%
0% 10% 20% 30% 40% 50% 60% 70% 80% 90%


What specific programming language do you use most often?
Python 54%
R 13%
SQL 8%
Java 6%
C/C++ 5%
C#/.NET 3%
Javascript/Typescript 3%
MATLAB 2%
SAS/STATA 1%
PHP 1%
Visual Basic/VBA 1%
Scala 1%
Bash .4%
Ruby .4%
Go .3%
Julia .1%
Other 1%
0%
10%
20%
30%
40%
50%
60%


What programming language would you recommend an aspiring data scientist to learn first?
Python
R
SQL
C++
MATLAB
Java
Scala
Javascript
SAS
VBA
Go
Other
None
75%
12%
5%
2%
1%
1%
.4%
.4%
.4%
.2%
.1%
.9%
.7%
0%
10%
20%
30%
40%
50%
60%
70%
80%

- 目前中国大部分教师和学生使用的诸如 SAS、SPSS、Stata、Eviews、Matlab、及微软的 Excel、word 等 Office 软件都是盗版的. 傻瓜软件使人弱智, 盗版软件危害无穷
- 中国大部分公立机构、学校不但在行政中使用美国商业 Office 软件, 甚至连雇员的上交材料也要求用美国商业软件 word 格式及 Excel 表格! 这是合法的吗?
- 这些公司不怕你盗版, 就怕你不上瘾. 中国自己的软件完全被这些盗版软件所遏制, 目前中国软件没有市场和用户基本上无法发展和生存.
- 使用盗版商业傻瓜软件既弱智又非法, 把国家安全抛到脑后. 看着华为的经历, 长点智慧吧!