UNIVERSITEIT GENT

FACULTY OF **ECONOMICS AND BUSINESS ADMINISTRATION**

Advanced Predictive Analytics

2016

Prof. Dr. Dirk Van den Poel

UNIVERSITEIT GENT

FACULTY OF **ECONOMICS AND BUSINESS ADMINISTRATION**

Introduction to R

Outline

1. Introduction
2. Why R?
3. R vs. SAS
4. Objective of this course

1. Introduction

- Teacher : Prof. Dr. Dirk Van den Poel
- Teaching assistants:
  - Matthias Bogaert
  - Steven Hoornaert
- Questions?
  - Help your fellow students and use the Minerva forum!
  - Mail: matthias.bogaert@ugent.be and steven.hoornaert@ugent.be
  - Appointment? Send an e-mail first!

1. Introduction

- Software:
  - R base: http://cran.r-project.org/
  - R Studio: http://www.rstudio.com/

1. Introduction

- Purpose of this session: go over the skills we require you to have in R before the start of *Advanced Predictive Analytics* (9 Feb 2016).
- The material given in this session, will not be revisited as such during the course. We assume that you are proficient in R before the start of the first course of *Advanced Predictive Analytics*.

1. Introduction

- Extra training material is available:
  - Coursera course: ‘R programming’
  - https://www.coursera.org/course/rprog
  - A lot of other interesting courses in the ‘Data Science’-specialization.

2. Why R?

- Strengths
  - Open-source: a lot of packages available (especially for predictive analytics)
  - Computing environment & low-level programming language
  - Efficient & fast syntax writing
  - Great for plotting
  - Matrix computing
- Weaknesses
  - Limited scalability (works on system memory)
    - Can be problematic when working with very large datasets!
  - Data manipulation is not always easy.
  - **Steep learning curve!**

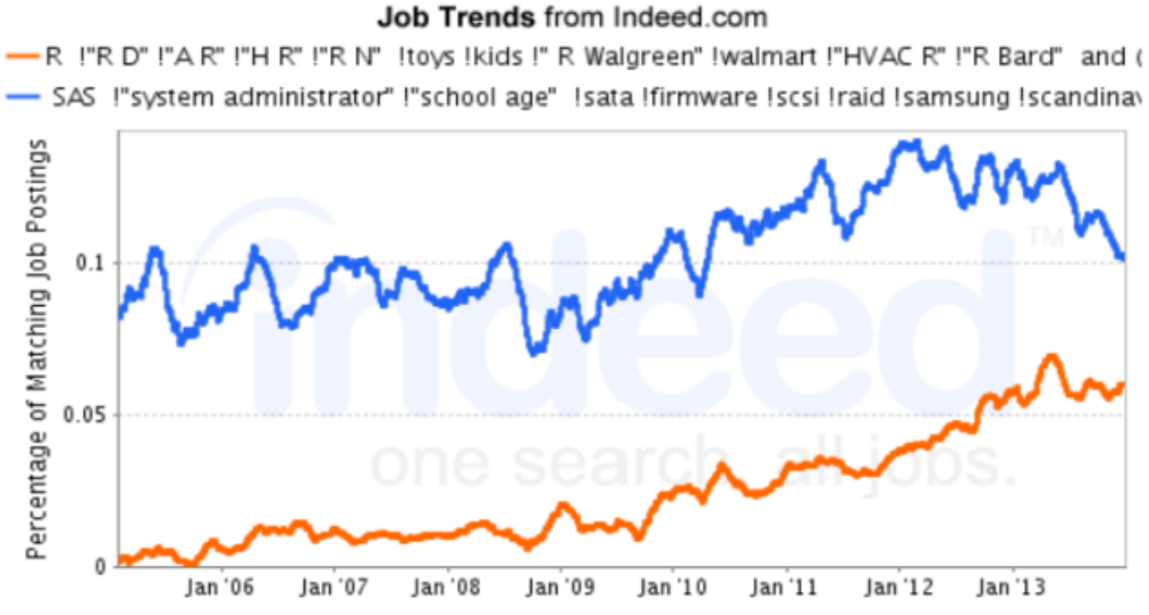
Why is R hard to learn?

- Too many commands
- Inconsistent syntax
- Poor ability to select variables
- Inconsistent in analyzing multiple variables
- Too many ways to transform & select variables
- Loop-a-phobia

Full article: http://r4stats.com/articles/why-r-is-hard-to-learn/

So why R?

Generally speaking, the pros outweigh the cons. Nowadays, R is gaining **more and more importance** in today's business world in favor of SAS. Furthermore, R is perfectly suited for modeling purposes and contains a lot of packages for predictive and prescriptive analytics. Hence R is considered the 'lingua franca' of statistics.



3. R vs SAS

- SAS
  - DATA step in SAS works with a built-in loop. SAS reads one row of data, evaluates the code line-by-line, executes and goes to the next row.
- R
  - R doesn’t have a built-in loop. R applies functions to columns i.e. R works in a vector-logic.

SAS

VIEWTABLE: Work.Data

	VAR1	VAR2
1	1	2
2	3	4
3	5	6

```
DATA data_out;
    SET data;
    VAR3=VAR1+VAR2;
RUN;
```

VIEWTABLE: Work.Data_out

	VAR1	VAR2	VAR3
1	1	2	3
2	3	4	7
3	5	6	11

VIEWTABLE: Work.Data

	VAR1	VAR2
1	1	2
2	3	4
3	5	6

```
DATA data_out;
    SET data;
    VAR3=VAR1+VAR2;
RUN;
```

VIEWTABLE: Work.Data_out

	VAR1	VAR2	VAR3
1	1	2	3
2	3	4	7
3	5	6	11

R

```
> data
  VAR1 VAR2
1    1    2
2    3    4
3    5    6
```

```
> data.frame(data,VAR3=data$VAR1+data$VAR2)
  VAR1 VAR2 VAR3
1    1    2    3
2    3    4    7
3    5    6   11
```

Conclusion: SAS vs R

- SAS
  - The basic building blocks are ROWS
- R
  - The basic building blocks are COLUMNS

R vs SAS vs Python

WINNER

Parameter	SAS	R	Python
Availability / Cost	2	5	5
Ease of learning	4.5	2.5	3.5
Data handling capabilities	4	4	4
Graphical capabilities	3	4.5	4
Advancements in tool	4	4.5	4
Job scenario	4.5	3.5	2.5
Customer service support and Community	4	3.5	3

Full article: http://www.analyticsvidhya.com/blog/2014/03/sas-vs-vs-python-tool-learn/

4. Objectives

- Data understanding
  - Basic data structures, data read-in, data exploration.
- Data preparation
  - Manipulate and transform data; combine into a basetable
    - *'The basetable is the beating heart of predictive modeling.'*

5. Basetable

- What is a *basetable*?
  - Every row = unique observation
    - e.g. Customer ID
  - Columns
    - Dependent variable: *what you want to predict*
      - E.g. churn (binary), cross-sell (binary or multi-label), CLV (numeric), …
    - Independent variables (or *predictors*): *data characteristics*
      - E.g. recency, frequency and monetary value

5. Basetable

Churn	Recency	Frequency	Monetary Value	Gender
1	20	50	300	1
1	36	12	600	0
0	5	5	200	0
1	45	1	50	1
0	5	60	1000	1

5. Basetable

- How to create a basetable?
  - Often different tables are involved
    - Ask yourself: *What do you want to model? What is your unique ID? How are the tables connected?*
      - An entity-relationship diagram (ERD) is informative.
  - Process every table seperately.
    - Create variables for every unique ID
      - A code flowchart can be informative.
  - Merge the different tables into one basetable
  - Handle missings after merging

In this session:

- DataUnd.R
- DataPrepPART1.R
- DataPrepPART2.R
- Extra_Exercises.R
- Assignment ‘Weibo’
  - Deadline : **9 Feb 2016, <u>before</u>** the start of the course
  - Hand in the assignment via dropbox (minerva)
  - Syntax: *StudentName_Assignment_Weibo.R*