

数据分析与处理技术

概述与工具准备

南京审计大学商学院物流管理系

考核要求

- 考试成绩+平时成绩 (60%+40%)
- 考试方式: 闭卷 写代码
- 平时成绩: 40% \longrightarrow 作业30%+保留10%
- 到课要求: 旷课一次扣除平时成绩的10%; 两次50%; 三次100%

学习要求

- 不要满足于明白，要在计算机上做出结果
- 大胆试错，尝试新技术新方法（成绩奖励）
- 严禁抄作业

技术前瞻

- 数据技术正在重塑经济形态
- 数据技术不等于计算机技术
- 物流管理对数据技术非常敏感
- 数据分析的基础课程再次加长

课程内容概括

数据操作

变量操作

图形语法

词法作用域

可视化技术

script编程

大数据处理

探索性分析

可视化分析

预测决策

关联分析

数据处理工具	特点
Excel	所见即所得，简单统计分析功能，报表强大
SPSS	容纳经典统计模型，操作简单，逐步加强编程能力
STATA/EVIEWS	计量经济工具，编程能力强大，编程语法易学
SAS	商务型分析工具，强大的数据管理和决策支持能力
LINGO	规划与概率问题计算工具
MATLAB/OCTIVE,SCILAB	矩阵式计算平台，界面友好，编程能力强大
MATHEMATICS/MAPLE	符号运算能力强大，文本与编程融合
C	编译型，适用面广，计算速度极快，代码较为繁琐
JAVA	面向对象，适用面广，代码较繁琐但功能独特
R	开源平台，统一的工具包管理，专用于数据分析
PYTHON	胶水语言，开源平台，大量框架支持
Julia	吸收R/Python的可扩展性，同时具备C/C++的速度

课程工具的配置

- 课程可用的工具主要有
 - R
 - Python
 - MATLAB
- 辅助工具
 - Github

<https://www.codecademy.com/>



统计学

package发布体系

CRAN: <https://cloud.r-project.org/>

Github: <https://github.com/>

学术体系

R Journal:
<https://journal.r-project.org/>

其他学术期刊:

JSS :<https://www.jstatsoft.org/index>

社区体系

Stackoverflow

Github

统计之都

GNU团队支持

IDE: <https://www.rstudio.com/>

框架:

tensorflow

Keras

Mxnet

计算机技术

网络参考资料表单

1. blog社区: 英文 StackOverFlow: <https://stackoverflow.com/>
Rblogger: <https://www.r-bloggers.com/>
中文 统计之都 <https://cosx.org/>
CSDN: <http://www.csdn.net/>
经管之家 <http://bbs.pinggu.org>
2. 网络电子书:
R project上的书
<https://www.r-project.org/doc/bib/R-books.html>
《Advanced R》
https://en.wikibooks.org/wiki/R_Programming
《Programming in R》
http://zoonek2.free.fr/UNIX/48_R/02.html#5
《Forecasting:Principle and Practice》
<https://otexts.org/fpp2/>

R programming环境配置

- 准备工具（开源工具）

- R语言

<https://www.r-project.org/>

- IDE

<https://www.rstudio.com/>

- 任务：安装R+Rstudio



工具配置

- 设置工作目录

- `setwd("c:\\Users\\Documents")`
- `setwd("c:/Users/Documents")`

由于单斜线在r中是转义功能，路径中的斜线需用双斜线或反斜线代替

- 工作空间(workspace)

- `ls()`函数：返回空间中所有对象
- `rm()`函数：删除空间中某对象
- `rm(ls())`

- 保存与调取空间

- save系列函数 `save.image()` `savehistory()`



```
save.image(file="myworkspace.RDATA")
```

- 与save函数对应的是load函数，调取存在硬盘上的记录文件

配置环境

- 环境参数
 - `options()` `options(digits=3)`
- 提示符号
 - `>` 待输入提示符
 - `+` 续行符
- 帮助命令
 - `?` 或 `help()` 调取某函数的说明信息
 - `#` 注释符

```
options(prompt='-')
```

Tips: R中单双引号作用相同

认识工作环境

- Console: 命令行环境
- Environment: 监控变量状态, 常用全局环境
- Script: 脚本文件, 或称程序文件
- Package: 包, 或称工具包, 动态加载专用的函数工具
- Rmarkdown: 文本与命令环境融合的书写文件, 以 markdown 为文本格式基础

输入记录：记录console中输入过的命令

```
Console Terminal x
~/

R version 3.4.4 (2018-03-15) -- "Someone to Lean On"
Copyright (C) 2018 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin15.6.0 (64-bit)

R是自由软件，不带任何担保。
在某些条件下你可以将其自由散布。
用'license()'或'licence()'来看散布的详细条件。

R是个合作计划，有许多人为之做出了贡献。
用'contributors()'来看合作者的详细情况
用'citation()'会告诉你如何在出版物中正确地引用R或R程序包。

用'demo()'来看一些示范程序，用'help()'来阅读在线帮助文件，或
用'help.start()'通过HTML浏览器来看帮助文件。
用'q()'退出R。

[Workspace loaded from ~/.RData]

> |
```

console:命令输入环境

Environment History Connections

Import Dataset

Global Environment

Environment is empty

Files Plots Packages Help Viewer

New Folder Delete Rename More

Home

	Name	Size	Modified
<input type="checkbox"/>	.RData	2.5 KB	Aug 28, 2018, 5:40 PM
<input type="checkbox"/>	.Rhistory	8.5 KB	Aug 28, 2018, 5:40 PM
<input type="checkbox"/>	anaconda2		
<input type="checkbox"/>	Applications		
<input type="checkbox"/>	Applications (Parallels)		
<input type="checkbox"/>	Creative Cloud Files		
<input type="checkbox"/>	Desktop		
<input type="checkbox"/>	Documents		
<input type="checkbox"/>	Downloads		
<input type="checkbox"/>	Env1.RDATA	4.4 MB	Aug 28, 2018, 5:39 PM
<input type="checkbox"/>	GitHub		
<input type="checkbox"/>	greyforecasting.bib	250.8 KB	Apr 14, 2018, 6:02 PM
<input type="checkbox"/>	greyforecasting.bib.bak	250.8 KB	Mar 27, 2018, 6:35 PM
<input type="checkbox"/>	greyforecasting.bib.sav	244.9 KB	Apr 15, 2018, 2:07 PM
<input type="checkbox"/>	Library		
<input type="checkbox"/>	Movies		

右下集合了工作区目录、绘图区和包管理几个常用功能

尝试基本运算操作

1.尝试基本运算

```
> 1+2  
[1] 3  
>
```

2.计算结果装入变量

```
> a<-1+2  
> |
```

3.也可以先将数据装入变量，进而取代操作数据

```
> a=1  
> b=2  
> y=a+b  
> y  
[1] 3  
>
```

4.将多个数据组合成向量，再装入变量a中

```
> a<-c(1,2,5,6)  
> a  
[1] 1 2 5 6
```

5.许多常见的数学函数可以直接使用，而且大部分支持向量化运算

```
> sin(5)  
[1] -0.9589243  
> a  
[1] 1 2 5 6  
> sin(a)  
[1] 0.8414710 0.9092974 -0.9589243 -0.2794155
```

Environment:监视变量和函数的状态

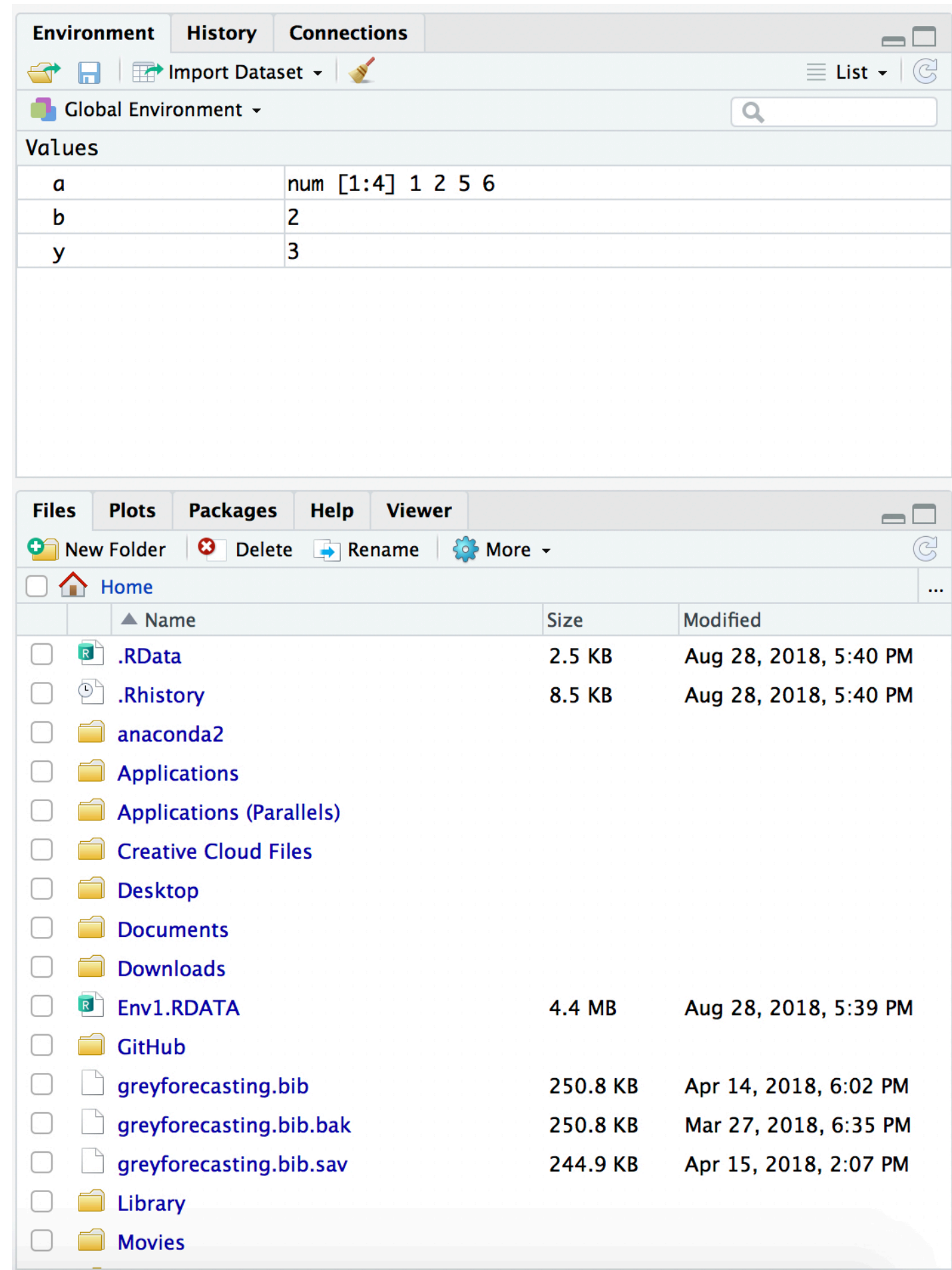
History:记录输入过的命令，与Env
一同存在workspace文件中

Files:工作目录下的文件状态

Plots:输出数据可视化图形的位置，
可以外置

Packages:监控和操作已安装的包，
简化包管理

Help:根据命令输出帮助文件信息





在线数据竞赛网的R代码环境

- Kaggle上的notebooke
- 启动云GPU的计算方式

<https://www.kaggle.com/>

Kernel->New Kernel->Edit Notebook

Competition->Fork->Edit Notebook




Test of R program

Draft saved R [Commit](#)

```
[ ]: library(tidyverse)
      train<-read_csv("../input/train.csv")
      summary(train)
```

Hide **Input** **Output** **Markdown** **Code**



```
dim(train)
train[1,]
```

```
[ ]: library(dplyr)
      g<-group_by(mtcars,cyl)
      summarise(g,var1=mean(dis))
```

```
[ ]: summary(mtcars)
      plot(mtcars$mpg,mtcars$disp)
```

Sessions

Interactive Session 0m:0s / 6h
CPU 0% RAM 157.9MB/17.2GB
GPU Off Disk 279.1MB/5.2GB

Versions

1 uncommitted draft
 Khadgar's draft based on V1

1 committed version
 V1 3mo +2 -0

Draft Environment

[+ Add Data](#)
 input (read-only)
 > House Prices: Advanced Regression T

Settings

Sharing Private, 0 collaborators

Language R

Docker exoplanetx/github.com_exopla...

GPU **BETA** GPU off

Internet **BETA** Internet blocked



Console

CPU 0% GPU OFF RAM 157.9MB/17.2GB Disk 279.1MB/5.2GB