Story of grading the achievements of the 1st IODS / Kimmo Vehkalahti

DATE / Sunday April 02 2017 08:23:30 Week=13 Day=092 I started by downloading (from MOOC Grade book) the Excel (gourmet) shit: INDEX *.xlsx / SINCE=TODAY xlsx=/OPEN /OPEN "OpenDataScience2017 Grades 20170402.xlsx" / here it is... :) 24425 02.04.17 08:21 First of all, read the data into R data frame and start checking. (Excel is OK for moving data, but it's not my tool for serious work.) R may be called from Survo R (that I here use) in various ways. The choice may depend, e.g., on what is most handy (or most fun!) at one time: 1) writing a prompt "R>" for the command 2) putting 'R' into the "control column" of the row 3) with R command referring to lines of edit field As an example (of having fun), let us use all these: R> library(openxlsx) # needed for reading Excel spreadsheets Rgrades <- read.xlsx(xlsxFile = "OpenDataScience2017 Grades 20170402.xlsx", colNames=TRUE) R CUR+1,CUR+2,END+2 / CUR = current line, END = last non-empty line of the edit field dim(grades) str(grades) The last one (option 3) reflects both the R input and output in the Survo editor: > dim(grades) [1] 130 24 > str(grades) 'data.frame': 130 obs. of 24 variables: \$ First.name : chr ... (hidden) \$ Surname : chr ... (hidden) \$ ID.number : num NA NA NA NA NA \$ Institution : chr NA NA NA NA .. : chr NA NA NA NA .. \$ Department \$ Email.address : chr ... (hidden) : chr "98.27" "45.68 \$ Course.total.(Real) : chr "20" "-" "11" \$ Workshop:.RStudio.Exercise.1.(submission).(Real) : chr "20" "-" "16" \$ Workshop:.RStudio.Exercise.2.(submission).(Real) : chr "20" "-" "7" " \$ Workshop:.RStudio.Exercise.3.(submission).(Real) : chr "20" "-" "13" \$ Workshop:.RStudio.Exercise.4.(submission).(Real) : chr "20" "-" "13" \$ Workshop:.RStudio.Exercise.5.(submission).(Real) : chr "5" "-" "5" "-\$ Workshop:.RStudio.Exercise.1.(assessment).(Real) : chr "5" "-" "5" "-\$ Workshop:.RStudio.Exercise.2.(assessment).(Real) : chr "5" "-" "5" "-\$ Workshop:.RStudio.Exercise.3.(assessment).(Real) : chr "5" "-" "5" "-\$ Workshop:.RStudio.Exercise.4.(assessment).(Real) : chr "5" "-" "5" "-\$ Workshop:.RStudio.Exercise.5.(assessment).(Real) : chr "97.69" "0" "1 \$ External.tool:.DataCamp:.Regression.and.model.validation.(Real) : chr "93.1" "-" "10 \$ External.tool:.DataCamp:.Logistic.regression.(Real) : chr "100" "-" "100 \$ External.tool:.DataCamp:.Clustering.and.classification.(Real) \$ External.tool:.DataCamp:.Dimensionality.reduction.techniques.(Real): chr "100" "-" "90. : chr "98.38" "91.35 \$ External.tool:.DataCamp:.R.Short.and.Sweet.(Real) : chr "-" "-" "-" "-\$ Assignment:.Final.assignment.(Real) : chr "1488516113" " \$ Last.downloaded.from.this.course

We (at least you!) could well continue with R functions learned on this course. However, I will continue this with Survo, because it is my strong comfort zone: I have used Survo for 25+ years and I won't abandon it, because I simply love to work with it! At the same time, I'm quite excited about the fancy things related to R that we have learned on this course. I got some howework, too! :D However, here the aim is to form the grades for you, and it is not the optimal situation for me to learn new tricks, as it would take more time than is available. With Survo I can fly much faster. :) So, I transfer the R data frame (grades) to a Survo data file: (no need to use CSV or txt files, Survo & R are deeply connected) FILE SAVE R>grades TO NEW GRADES FILE SHOW GRADES / I always like to browse the data files! In this case, Survo gives a fair warning, because the names of the data do not conform to its strict policy (names not unique to the first 8 characters may cause trouble in subsequent operations): Saving R data frame grades to file GRADES: Field name Workshop appears at least twice! Field name External appears at least twice! Survo data file will be created, but use FILE STATUS/UPDATE for field name editing! Let us look at the structure of the Survo data file GRADES: (these variable "names" come automatically from the MOOC grade book) FILE STATUS GRADES / this command writes the current structure below: Copied from R data frame grades FIELDS: (active) 1 SA_ 17 First.name 2 SA_ 12 Surname 3 NA_ 8 ID.number 4 SA_ 22 Institution 5 SA_ 25 Department 6 SA_ 38 Email.address 18 Course.total.(Real) 7 SA_ 8 SA_ 2 Workshop:.RStudio.Exercise.1.(submission).(Real) 2 Workshop:.RStudio.Exercise.2.(submission).(Real) 9 SA_ 10 SA_ 2 Workshop:.RStudio.Exercise.3.(submission).(Real) 11 SA_ 2 Workshop:.RStudio.Exercise.4.(submission).(Real) 12 SA_ 2 Workshop:.RStudio.Exercise.5.(submission).(Real) 13 SA_ 1 Workshop:.RStudio.Exercise.1.(assessment).(Real) 14 SA_ 1 Workshop:.RStudio.Exercise.2.(assessment).(Real) 15 SA_ 1 Workshop:.RStudio.Exercise.3.(assessment).(Real) 16 SA_ 1 Workshop:.RStudio.Exercise.4.(assessment).(Real) 17 SA_ 1 Workshop:.RStudio.Exercise.5.(assessment).(Real) 18 SA_ 5 External.tool:.DataCamp:.Regression.and.model.validation.(Real) 19 SA_ 18 External.tool:.DataCamp:.Logistic.regression.(Real) 20 SA_ 18 External.tool:.DataCamp:.Clustering.and.classification.(Real) 21 SA_ 18 External.tool:.DataCamp:.Dimensionality.reduction.techniques.(Re 22 SA_ 5 External.tool:.DataCamp:.R.Short.and.Sweet.(Real) 23 SA_ 2 Assignment: .Final.assignment. (Real) 24 SA_ 10 Last.downloaded.from.this.course END Survo data file GRADES: record=307 bytes, M1=34 L=64 M=24 N=130 (N: numeric, S: string vars - both types may be used in computations, which is often handy: the content of the variable is primary, while its type is secondary - this is not typical in stats programs, however) Well, let us modify the names and the descriptions of the names so that there will be no trouble with the data integrity, and that we can more easily follow what is going on (data documentation is VERY important). I let a few names be as they are, as I won't need them ("NA!"), but

also change some, e.g., the email, which I need to use as the key for matching the contents with the DataCamp analytics data and Oodi student numbers data that I have formed earlier). Recall that the MOOC area is for MOOCs, i.e., it does not contain info about the student numbers. Of course, names are bad for keys, so the email is the only possibility.

In addition, I create new DC variables and some Oodi variables to be used in communicating the results to the registration admins of UH. (You don't need to know all this, I just think aloud while I work!) :) FILE EXPAND GRADES / some more space for new variables FILE UPDATE GRADES / this command updates the edited structure below Copied from R data frame grades FIELDS: (active) 1 SA_ 17 First.name (MOOC platform) 2 SA_ 12 Surname 8 ID.number (MOOC platform) 3 NA_ (NA!) 4 SA_ 22 Institution (NA!) 5 SA_ 25 Department (NA!) 6 SA_ 38 email Email.address (KEY for combining the data sets!) 7 SA_ 18 Course.total.(Real) (NA!) 8 SA_ 2 WS1s RStudiol: Tools and methods for open ... (submit) 9 SA_ 2 WS2s RStudio2: Regression and model validation (submit) 10 SA_ 2 WS3s RStudio3: Logistic regression (submit) 11 SA_ 2 WS4s RStudio4: Clustering and classification (submit) 12 SA_ 2 WS5s RStudio5: Dimensionality reduction techniques (submit) 13 SA_ 1 WS1a RStudio1: Tools and methods for open ... (assess) 14 SA_ 1 WS2a RStudio2: Regression and model validation (assess) 15 SA_ 1 WS3a RStudio3: Logistic regression (assess) 16 SA_ 1 WS4a RStudio4: Clustering and classification (assess) 17 SA_ 1 WS5a RStudio5: Dimensionality reduction techniques (assess) 18 SA_ 5 dc2 DataCamp: Regression and model validation (MOOC/LTI) 19 SA_ 18 dc3 DataCamp: Logistic regression (MOOC/LTI) 20 SA_ 18 dc4 DataCamp: Clustering and classification (MOOC/LTI) 21 SA_ 5 dc1 DataCamp: Regressionality reduction techniques (MOOC/LTI) 22 SA_ 5 dc1 DataCamp: R Short and Sweet (extra!) (MOOC/LTI) 23 SA_ 2 finalAss Final Assignment (raw score) 7 SA_ 18 Course.total.(Real) (NA!) 23 SA_ 2 finalAss Final Assignment (raw score) 24 SA_ 10 Last.downloaded.from.this.course (NA!) 25 NA_ 4 DC2 DataCamp: Regression and model validation (DC Excel) 4 DC3 DataCamp: Logistic regression (DC Excel) 26 NA_ 4 DC3 DataCamp: Dogistic regression (so Encer,
4 DC4 DataCamp: Clustering and classification (DC Excel)
4 DC5 DataCamp: Dimensionality reduction techniques (DC Excel)
4 DC1 DataCamp: R Short and Sweet (extra!) (DC Excel, total) 27 NA_ 28 NA_ 29 NA_ 30 SA_ 17 etunimi First name (Oodi) 31 SA_ 12 sukunimi Surname (Oodi) 32 SA_ 9 opnro Student number (Oodi) 33 NA_ 2 FinalAss Final Assignment (scaled) END Survo data file GRADES: record=305 bytes, M1=34 L=64 M=24 N=130

FYI: There was something wrong with the LTI integration of the MOOC platform and the DataCamp platform, as some scores were not right on the MOOC. Thanks for reporting those errors to me already during the course! Finally, I decided to forget about the MOOC scores altogether and use the data exported from the DataCamp. I will report these problems to DataCamp, but for now we can manage perfectly, no worries! The only nuisance is that the scores that everyone can see on the MOOC area are quite useless. Do not waste your time with them! :) In this GRADES data I will have everything in good order, as you may notice.

In passing I combine the information from other datasets I prepared earlier: FILE SHOW DCGRADES / full scores if done before deadlines, 90% otherwise! Snapshot from another Survo edit field, where I completed the DC scores (2-5):

Assignment giving days (1) and deadlines for completion (2): ready-made block to copy to DATEs below: 1DATE 26.01.2017, Julian / Thu Jan 26 2017 Julian_day=2457780 2DATE 01.02.2017, Julian / Wed Feb 01 2017 Julian_day=2457786 1DATE 02.02.2017, Julian / Thu Feb 02 2017 Julian_day=2457787 2DATE 08.02.2017, Julian / Wed Feb 08 2017 Julian_day=2457793 1DATE 09.02.2017, Julian / Thu Feb 09 2017 Julian_day=2457794 2DATE 15.02.2017, Julian / Wed Feb 15 2017 Julian_day=2457800 1DATE 16.02.2017, Julian / Thu Feb 16 2017 Julian_day=2457801 2DATE 22.02.2017, Julian / Wed Feb 22 2017 Julian_day=2457807 FILE SHOW DATACAMP2 FILE EXPAND DATACAMP2 For all DATE commands below, the format is IDATE=MM/DD/YYYY (DC: US style!) VAR c01time1:2=MISSING TO DATACAMP2 VAR c01time2:2=MISSING TO DATACAMP2 DATE DATACAMP2 / VARS=c0lstart(D),c0ltime1(J) JULIAN_DAY0=2457780 DATE DATACAMP2 / VARS=c0lended(D),c0ltime2(J) JULIAN_DAY0=2457786 VAR c02time1:2=MISSING TO DATACAMP2 VAR c02time2:2=MISSING TO DATACAMP2 DATE DATACAMP2 / VARS=c02start(D),c02time1(J) JULIAN_DAY0=2457787 DATE DATACAMP2 / VARS=c02ended(D),c02time2(J) JULIAN_DAY0=2457793 VAR c03time1:2=MISSING TO DATACAMP2 VAR c03time2:2=MISSING TO DATACAMP2 DATE DATACAMP2 / VARS=c03start(D),c03time1(J) JULIAN DAY0=2457794 DATE DATACAMP2 / VARS=c03ended(D),c03time2(J) JULIAN_DAY0=2457800 VAR c04time1:2=MISSING TO DATACAMP2 VAR c04time2:2=MISSING TO DATACAMP2 DATE DATACAMP2 / VARS=c04start(D),c04time1(J) JULIAN_DAY0=2457801 DATE DATACAMP2 / VARS=c04ended(D),c04time2(J) JULIAN_DAY0=2457807 FILE SHOW DATACAMP2 7 / looks great! most students followed the schedules. (I did that on Sunday morning.) Now, back to this business: FILE COPY DCGRADES TO GRADES / MATCH=email MODE=2 VARS=DC2,DC3,DC4,DC5,DC1 FILE SHOW GRADES / FIELD=DC2 Student numbers retrieved already in mid-February (luckily! took some time...) FILE COPY OPNRO1 TO GRADES / MATCH=email MODE=2 VARS=opnro,etunimi,sukunimi FILE SHOW GRADES / FIELD=opnro Now, everything is in the same data file. Great. Awesome. Magnificent. Simple. Also, prepare for scaling of the various scores: (see below) VAR FinalAss=5*finalAss TO GRADES / scaling from [0,40] to [0,200] FILE SHOW GRADES / FIELD=FinalAss (Variables names are case-sensitive, similarly as in R - that is smart!)

Some simple summaries follow (boring statistics removed by LINEDELS): LINEDEL CUR+1 END "min=" LINEDEL CUR+2 END "max=" LINEDEL CUR+3 END "mean=" LINEDEL CUR+4 END "lower " #STAT GRADES CUR+2 / CLASSMAX=9 (compact automatic classification requested) Basic statistics: GRADES N=130 Variable: WS1s RStudio1: Tools and methods for open ... (submit) f % class width=5 up.limit 0 0 0.0 5 0 0.0 10 6 4.6 ***** 15 20 RStudio2: Regression and model validation (submit) Variable: WS2s up.limit f % class width=5 0 3 2.3 *** 5 10 2 1.5 ** 9.2 ********* 15 12 20 RStudio3: Logistic regression (submit) Variable: WS3s up.limit f % class width=5 0 1 0.8 * 5 2.3 *** 4.6 ***** 3 6 10 15 Variable: WS4s RStudio4: Clustering and classification (submit) up.limit f % class width=5 0 5 0 0.0 6 4.6 ***** 10 13 10.0 *********** 15 20 Variable: WS5s RStudio5: Dimensionality reduction techniques (submit) up.limit f % class width=5 0 5 3.8 **** 5 1 10 0.8 * 13 10.0 *********** 15 2.0 Variable: WS1a RStudiol: Tools and methods for open ... (assess) WS1a f % 0 5 Variable: WS2a RStudio2: Regression and model validation (assess) f WS2a % 0 5 Variable: WS3a RStudio3: Logistic regression (assess) f WS3a % 0 5 Variable: WS4a RStudio4: Clustering and classification (assess) WS4a f 8 0 5

RStudio5: Dimensionality reduction techniques (assess) Variable: WS5a f % WS5a Ω 5 Variable: DC2 DataCamp: Regression and model validation (DC Excel) N(missing)=13 f % class width=20 up.limit б.8 ****** 0 8 1.7 ** 2 20 40 0 0.0 5.1 ***** 6 60 4 3.4 **** 80 100 Variable: DC3 DataCamp: Logistic regression (DC Excel) N(missing)=13 f % class width=20 up.limit 18 15.4 ************** 0 20 3 2.6 *** 40 2 1.7 ** 1 0.9 * 60 3.4 **** 80 4 100 Variable: DC4 DataCamp: Clustering and classification (DC Excel) N(missing)=13 up.limit f % class width=20 22 18.8 ***************** 0 1 0.9 * 20 2 40 1.7 ** 1 2 0.9 * 1.7 ** 60 80 100 Variable: DC5 DataCamp: Dimensionality reduction techniques (DC Excel) N(missing)=13 up.limit f % class width=20 28 23.9 **************************** 0 0 0.0 20 40 1 0.9 * 2 1.7 ** 60 5.1 ***** б 80 Variable: DC1 DataCamp: R Short and Sweet (extra!) (DC Excel, total) N(missing)=13 up.limit f % class width=50 5 4.3 **** 0 50 3 2.6 *** 100 3 2.6 *** 150 3 2.6 *** 200 Variable: FinalAss Final Assignment (scaled) up.limit f % class width=50 0 3.8 ***** 50 5 9.2 ********* 100 12 21 16.2 **************** 150 41 200

Quick draft of my scaling idea on Sunday: (in order to sum them up) (thanks to Tuomo for a quick chat in Slack for the final version!)

workshop 1-5 á [0,20] -> scale to [0,200] (finally: WS1 only 100)
peer-reviews á [0,5] -> scale to [0,50] (finally: 40 each)
DataCamp 2-5 á [0,100] (OK as such)
DataCamp 1 [0,200] <- too much? (R Short and Sweet) (finally: 100)
FinalAss [0,200] -> re-scale to [0,400] - in the future, the Final
Assignment will be obligatory (conclusion!)

MAX SCORE would be 1*100+4*200+5*40+4*100+100+400=2000 (nice!) - so, proceed:

VAR TOTAL=RStudio+PeerReview+DataCamp+RShortAndSweet+FinalAssignment TO GRADES

with the formulas of this "symbolic computation" (editorial arithmetics) as:

RStudio=5*WS1s+10*WS2s+10*WS3s+10*WS4s+10*WS5s PeerReview=8*WS1a+8*WS2a+8*WS3a+8*WS4a+8*WS5a (dropped from 10 to 8 to get 2000) DataCamp=DC2+DC3+DC4+DC5 RShortAndSweet=DC1/2 (DC1 was scaled to [0,200] earlier: too much!) FinalAssignment=2*FinalAss (I wanted to give this a proper weight - great jobs!)

Historical remark (or two) for your information:

This free-form, interactive way of doing and documenting statistical analyses and other tasks, was invented (quite accidentally!) by prof. Seppo Mustonen in 1979, in relation to printing music notation (!) for his son Olli Mustonen: http://survo.fi/demos/#ex6

That was the origin of **Survo Editor** that has existed for 2017-1979=38 years, on different platforms and operating systems, implemented in various ways. It is still going strong, currently in a form of an R package "muste", see http://survo.fi/muste/

Survo - the life-work of Seppo - started already in the beginning of 1960s: http://survo.fi/presentation/history.html Now it's open-source - who would have guessed that in 1960s/-70s/-80s/-90s??

For a comprehensive **list of publications** related to Survo, take a look at http://www.survo.fi/publications/ Especially, the years 1976-1982 are interesting in shedding light to the early stages of **SURVO 76**, one of the first interactive stats programs in the world, in the middle of the era of computing centres and mainframes.

In 1985, the C language version of Survo - SURVO 84C - was started, which is also interesting, as it still forms the code base of the current versions, SURVO MM and Survo R. Programmers, look at http://www.helsinki.fi/survo/c/

Another remark for your information - the background of why we are now here:

Until 2016, we had a basic course of Data Analysis with Survo (*), but last summer I decided to give it up, and make room for a new course - Intro Open Data Science.

That was not an easy decision for me, but I knew that it was time to "change the course" - the way I tend to do, cf. the other course I have radically changed: https://goo.gl/mEB8BX (ugly URL from TUHAT shortened by goo.gl)

(*) including a bit & byte of **SPSS** (for historical reasons).

OK, back to this! Thx for your attention :) What do we have in the TOTAL score?

STAT GRADES CUR+2 / VARS=TOTAL CLASSMAX=20

~RStudio+PeerReview+DataCamp+RShortAndSweet+FinalAssignment Variable: TOTAL N(missing)=13 in obs.#7 (first matching student's name removed from here) min=0 max=2000 in obs.#42 (student's name removed from here) - congrats to you! ;) mean=1297.796 stddev=679.0965 skewness=-0.709393 kurtosis=-1.032102 lower_Q=716.6667 median=1583.333 upper_Q=1856.25 up.limit f % class width=200 3 2.6 *** 0 7 6.0 ***** 200 0 _____ I like sketching ideas, doing 800 3 2.6 *** 1000 5 4.3 **** calculations, analyses etc. freely in the Survo editor. It's much like _____ using paper & pen!
 1200
 5
 4.3

 2
 1400
 7
 6.0

 (may look chaotic, I know...) 7 6.0 ****** _____ **3** 1600 12 10.3 ********** _____ _____ According to those dashed lines, let us form the grades as follows: CLASSIFY GRADES, Grades, TOTAL, GRADE:1 (1 byte num var is enough) CLASSIFICATION Grades 1801 - 2000: 5 1601 - 1800: 4 1401 - 1600: 3 1001 - 1400: 2 601 - 1000: 1 OTHERS: 0 END And how does that look? Not bad! About 1/3 excellent, 1/3 did not finish (no time etc.) and 1/3 in between, with an increasing trend from 1 to 4: (*) STAT GRADES CUR+1 / VARS=GRADE Basic statistics: GRADES N=130 Variable: GRADE Ordinal scale f GRADE 00 0 8 6.2 ****** 1 9.2 *********** 9.2 ********** 2 12 3 12 13.1 *********** 17 4 5 (*) Recall: 5=excellent, 4=very good, 3=good, 2=satisfactory, 1=adequate A good way to ensure that everything is OK is to create a cross-table:

TAB GRADES END+2 / VARIABLES=GRADE, TOTAL

The classifications must be given somewhere, e.g., GRADE=0,0,1,2,3,4,5 and TOTAL=0,600(0000-0600),1000(0601-1000),1400(1001-1400),1600(1401-1600),& 1800(1601-1800),2000(1801-2000),MISSING([MISSING]) [a bit messy, but...]

TABLE	GRAD	ES1 A	,B,F	N=	130						
A		GRADE	0	1	2	3	4	5			
TOTAL		* * * * *									
0000-	0600		28	0	0	0	0	0			
0601-	1000		0	8	0	0	0	0			
1001-	1400		0	0	12	0	0	0			
1401-	1600		0	0	0	12	0	0			
1601-	1800		0	0	0	0	17	0			
1801-	2000		0	0	0	0	0	40			
B[MISS	ING]		13	0	0	0	0	0			
Chi s	quare	=650.	0 df=	30 P	=0.0	0000					
It mi COPY REPLA	A,B,C CE "	De eve CUR+3 0 " "	n bet "	C /	LINE	che z ES=Cl	JR+4,	are B	removed	altoget.	ner:
7			0	1	2	2	4	F			
A		GRADE	0	T	2	3	4	5			
101AL	0600		20								
0000-	1000		20	0							
1001-	1400			8	10						
1401-	1600				ΤZ	10					
1401-	1000					⊥Z	1 7				
1001-	T800						т/	10			
T80T-	2000							40			

To be honest, I spotted an error in my classifications! :) Shit happens, that's why it is so important to document one's work. I fixed that and sent a new message to the admins in order to raise 5 "1"s to "2"s. :)

Thanks for reading this story. It's been fun to study together with you!

We are looking forward to your precious thoughts and feedback:

https://elomake.helsinki.fi/lomakkeet/78366/lomake.html
(so far, only 18 of you have sent your ideas - come on, guys!)

DATE / Monday April 03 2017 12:07:12 Week=14 Day=093

B[MISSING] 13

. . .

DATE / Monday April 03 2017 19:36:28 Week=14 Day=093

Did not have time to send this until the evening! Now, got to go home, pack my stuff for the Estonian trip and have a power sleep!