Excel vs. R: the search for a better language in data analysis

senjiew

10/31/2017

Introduction

In this post, I will discuss some of the similarities and differences between R and Excel. It will cover topics such as ease of use, data manipulation, reproducibility, graphs, and pricing. I wanted to choose this topic because as someone who is interested in finance, Excel is an essential part of data analysis and that it is critical to perform various different functions including the financial statements, and regression analysis. However, ever since I have been exposed to R from this course, I feel like I have gained a better perspective on how to analyze csv data in a more text-based software. Even though it demands more from me in terms of knowing its syntax, it does gives me more control over data manipulation. Therefore, I wanted to do more research in both Excel and R to determine the right language for me going into a career in finance.

Ease of Use

R uses a text-based editor that requires the user to be at the minimum, literate in computer science. Its commands usually have to be typed out into a console using its syntax. For example:

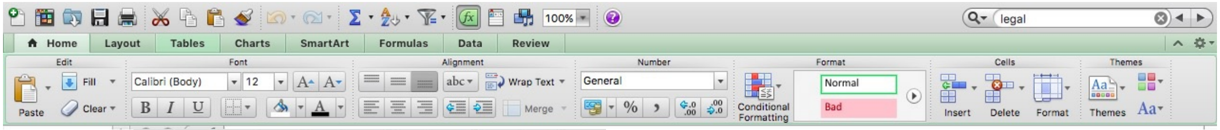
```
c("Hello World")
```

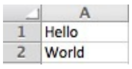
```
## [1] "Hello World"
```

```
matrix(1:25,nrow=5,ncol=5,byrow = TRUE)
```

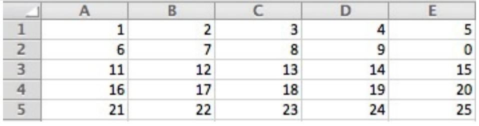
```
##      [,1] [,2] [,3] [,4] [,5]
## [1,]    1    2    3    4    5
## [2,]    6    7    8    9   10
## [3,]   11   12   13   14   15
## [4,]   16   17   18   19   20
## [5,]   21   22   23   24   25
```

On the other hand, Excel uses a Graphical User Interface (GUI) that allows users to click on different options to execute commands. Data can be inputted in individual cells and is on display at all times as it acts more like a data frame in R For example:





	A
1	Hello
2	World



	A	B	C	D	E
1	1	2	3	4	5
2	6	7	8	9	0
3	11	12	13	14	15
4	16	17	18	19	20
5	21	22	23	24	25

More resources regarding basic interactions with R can be found here while more resources regarding Excel's GUI including its macros can be found here.

Data Manipulation

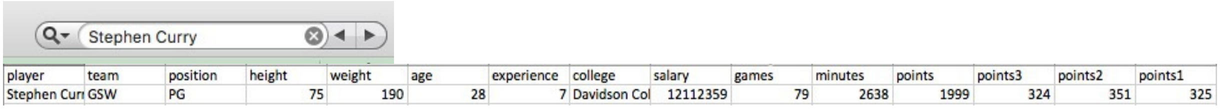
As both software are created for data manipulation and analysis, R has the abilities to manipulate different types of data that are not atomic using lists and data frames while Excel does not as it cannot differentiate data types within its cells.

```
dat <- read.csv("data/nba2017-players.csv",stringsAsFactors = FALSE)
filter(dat, player == "Stephen Curry")
```

```
##          player team position height weight age experience
## 1 Stephen Curry  GSW       PG     75    190  28          7
##            college   salary games minutes points points3 points2 points1
## 1 Davidson College 12112359    79    2638   1999     324     351     325
```

As evident, R can easily filter out a specific player and his stats with minimal lag even working on my laptop with only a 4GB RAM. The function dplyr is also very useful when manipulating data, given its ease of understanding and ease of arguments. A helpful tutorial of dplyr can be found here.

On Excel, the Find function becomes very important to look for players as the spreadsheet format displays all the data at the same time.

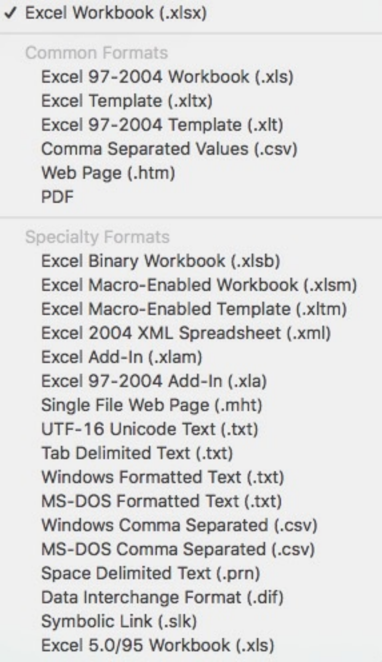


player	team	position	height	weight	age	experience	college	salary	games	minutes	points	points3	points2	points1
Stephen Curr	GSW	PG	75	190	28	7	Davidson Col	12112359	79	2638	1999	324	351	325

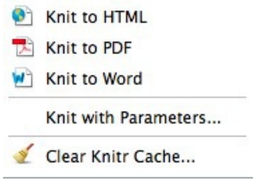
As mentioned previously, the Excel outputs are usually a lot more user friendly, as data can be displayed in individual cells. However, even with 441 players and their respective statistics, I experienced a lot of lag and a slow down in my computer. Furthermore, Excel did not displayed the data types of the columns such as "character" or "integer", which means that it assumes all data type to be the same. The lack of differentiation can be troublesome for data analysis, as it can slow down the analysis and the overall process.

Reproducibility

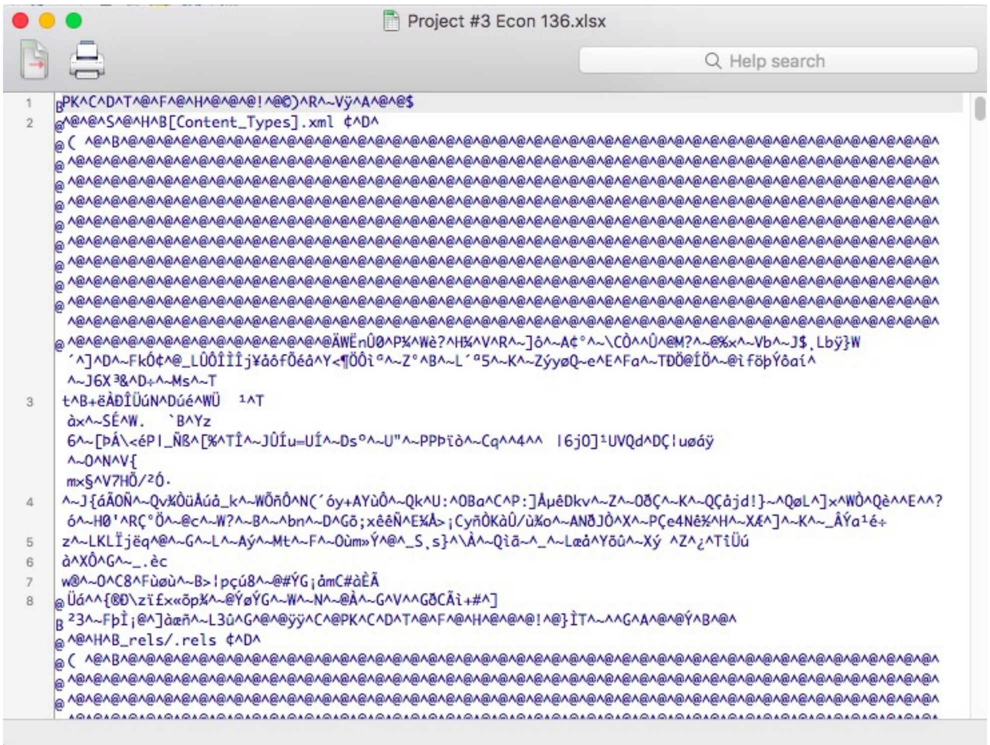
The reproducibility of R is unambiguously better than that of Excel. R has built in package knitr that allows it to be saved in many different types of documents including md, github, and html. Since all the formulas are coded into the documents, any text editing software is going to be able to read and reproduce the data. Excel on the other hand also can save most of its files in different formats including txt and csv, it deletes all the formulas used to format the workbook and calculate the numbers. Another common format saved in Excel is xlsx, the built in format that Microsoft Excel uses that allows formulas to be saved. However, this is not reproducible by other text editing software. This makes any data analyzed in Excel very difficult to reproduce.


✓ Excel Workbook (.xlsx)
Common Formats
Excel 97-2004 Workbook (.xls)
Excel Template (.xltx)
Excel 97-2004 Template (.xlt)
Comma Separated Values (.csv)
Web Page (.htm)
PDF
Specialty Formats
Excel Binary Workbook (.xlsb)
Excel Macro-Enabled Workbook (.xlsm)
Excel Macro-Enabled Template (.xltm)
Excel 2004 XML Spreadsheet (.xml)
Excel Add-In (.xlam)
Excel 97-2004 Add-In (.xla)
Single File Web Page (.mht)
UTF-16 Unicode Text (.txt)
Tab Delimited Text (.txt)
Windows Formatted Text (.txt)
MS-DOS Formatted Text (.txt)
Windows Comma Separated (.csv)
MS-DOS Comma Separated (.csv)
Space Delimited Text (.prn)
Data Interchange Format (.dif)
Symbolic Link (.slk)
Excel 5.0/95 Workbook (.xls)

Excel Formats



Knitr Formats



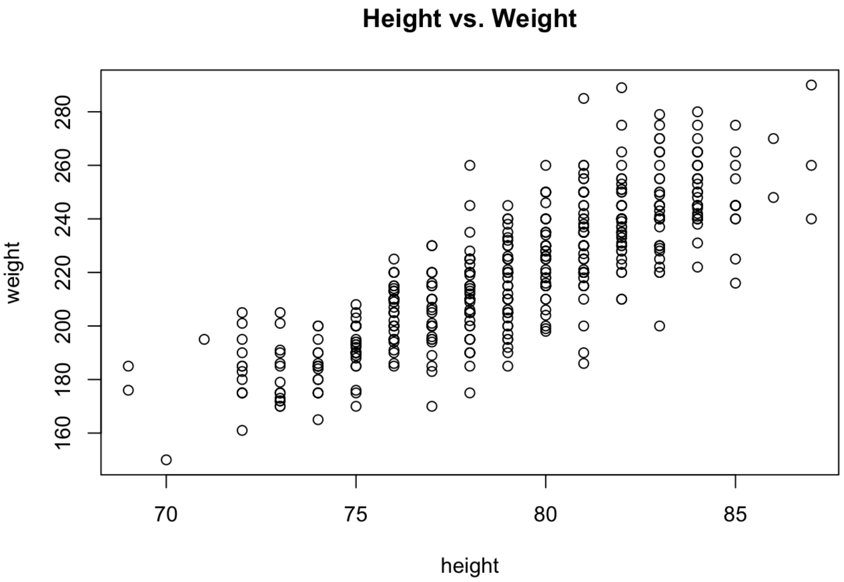
```
BPK^C^D^T^@^F^@^H^@^@^@!^@©)^R^~Vÿ^A^@^@$
@^@^@^S^@^H^B[Content_Types].xml ¢^D^
@( ^@^B^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@
@ ^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@
@ ^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@
@ ^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@
@ ^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@
@ ^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@
@ ^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@
@ ^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@
@ ^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@
  ^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@
@ ^@^@^@^@^@^@^@^@^@^@^@^@^@ÄWËnÛØ^P%^Wè?^H¾^V^R^~]ô^~A¢°^~\CÒ^^Û^@M?^~@%x^~Vb^~J$¸Lbÿ}W
  ´^]^D^~FkÓ¢^@_LÛÔÎÌÎj¥ȧôfÕéȧ^Y<¶ÖÔìª^~Z°^B^~L´ª5^~K^~ZýyøQ~e^E^Fa^~TÐÖ@ÍÖ^~@ìföþÝôaí^
  ^~J6X³&^D÷^~Ms^~T
 t^B+ëÀÐÎÜúN^Dúé^WÜ  ¹^T
  àx^~SÉ^W.  `B^Yz
  6^~[þÁ\<éP|_Ñß^[%^TÎ^~JÛÍu=UÍ^~Ds°^~U"^~PPþïò^~Cq^^4^^ |6j0]¹UVQd^DÇ¦uøáÿ
  ^~0^N^V{
  mx§^V7HÕ/²Ó·
 ^~J{áÄOÑ^~Qv¾ÒüÅúȧ_k^~WÕñÔ^N(´óy+AYùÔ^~Qk^U:^OBa^C^P:]ÅµêDkv^~Z^~OðÇ^~K^~QÇȧjd!}~^QøL^]x^WÒ^Qè^^E^^?
  ó^~HØ'^RÇ°Ö^~@c^~W?^~B^~^bn^~D^Gõ;xêêÑ^E¾Å>¡CyñÒKàÛ/ù¾o^~ANðJÒ^X^~PÇe4Nê½^H^~XÆ^]^~K^~_ÅÝa¹é÷
 z^~LKLÏjëq^@^~G^~L^~Aý^~Mt^~F^~Oùm»Ý^@^_S¸s}^\À^~Qìã~^_^~Lœȧ^Yõû^~Xý ^Z^¿^TìÜú
 à^XÔ^G^~_.èc
 w@^~0^C8^Fùøù^~B>¦pçú8^~@#ÝG¡ȧmC#àÈÄ
@Üá^^{®Ð\zï£x«õp¾^~@ÝøÝG^~W^~N^~@Å^~G^V^^GðCÃì+#^]
B²3^~FþÌ¡@^]àœñ^~L3û^G^@^@ÿÿ^C^@PK^C^D^T^@^F^@^H^@^@^@!^@}ÌT^~^^G^A^@^@Ý^B^@^
@^@^H^B_rels/.rels ¢^D^
@( ^@^B^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@
@ ^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@
@ ^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@^@
```

This is what happens when an xlsx file is opened up in a text editor such as R

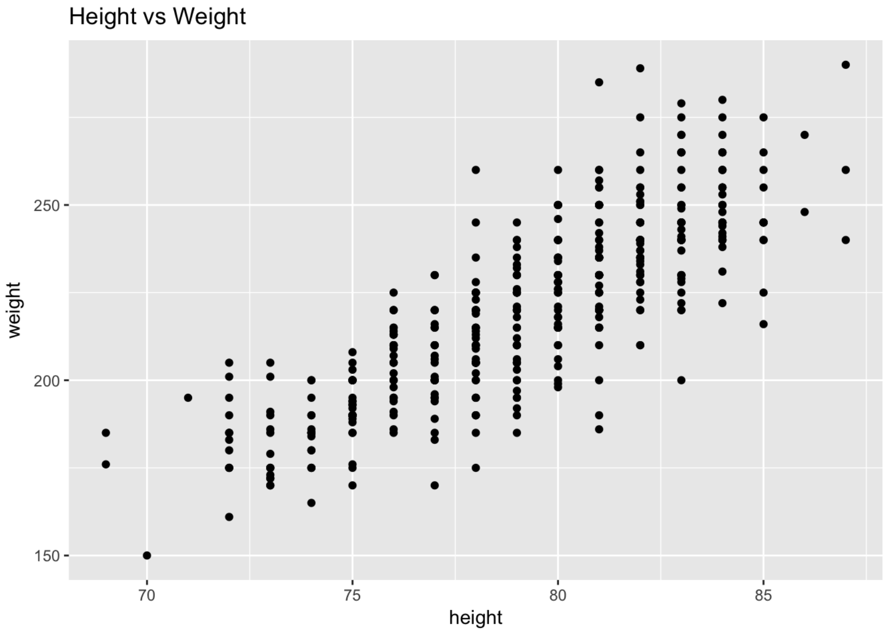
Graphs

R offers multiple ways to visually represent data, including but not limited to plot() and ggplot() in the tidyverse package. For example:

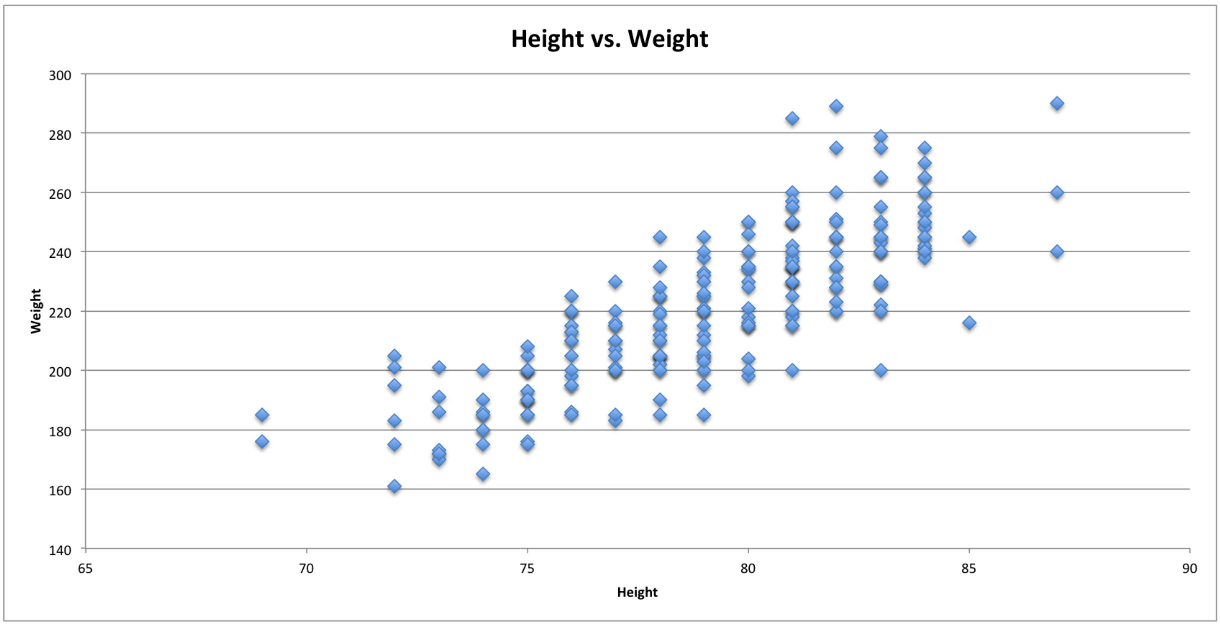
```
plot(dat$height, dat$weight, xlab = 'height', ylab = 'weight',main = "Height vs. Weight")
```



```
ggplot(data = dat, aes(height, weight)) +
  geom_point() +
  ggtitle("Height vs Weight")
```



On Excel:



Again, because Excel stores data as a singular data type, it falls short on speed. The amount of time it took me to make the first two graphs was significantly less than that of the third. Even though it might be easier to click on the options that Excel provides, I believe that the extra speed and performance that R provides exceeds Excel in terms of large data processing and thus the visual representation of such data.

Additionally, while Excel's graph seems like it looks better with the 3D dots that are colored in. There are a very limited amount of graphical parameters that can be applied to a graph. On the other hand, the graphical options available in packages such as ggplot2 far exceeds Excel's graph making capabilities in terms of aesthetics. To learn more about ggplot, here is a great place to start!

Pricing:

On top of their difference in capabilities, Excel and R differ significantly in pricing as well. Excel is a part of Microsoft Office, which can be purchased as a one-time order for $149 or on a subscription for $7 a month for 1 user or $10 a month for 5 users. The yearly subscription pricing offers a 16% discount to the monthly model at $70 a year and $100 dollars a year respectively Microsoft Office. On the other hand, R is completely free to download for personal use, but can cost more than $10,000 for a commercial license R Studio.

For someone who is new to data analysis, perhaps a monthly subscription in Microsoft Office is the way to go as it provides a variety of services outside of Excel including Word and PowerPoint, which are all important tools to use in the business world. However, the fact that R is free should also be considered when deciding the optimal software for each individual as the costs can stack up with Excel and that R indeed provides all the necessary tools Excel has, but just in a different way.

Conclusion

Therefore, while Excel provides excellent resources for people who are new to the world of data analysis with its easy -to-use GUI, R proves to be a more powerful software with more options and opportunities to explore. However, as Excel is still the golden software used by everyone in finance and that I am still a beginner in R. I will try to work on most of my data in Excel, but I will definitely try to gradually start to do more in R.

For more resources on R and Excel, a great place to start is Stack Overflow. It provides a platform for people all over the world to discuss coding related questions and it is completely free.

References

1. http://www.r-tutor.com/r-introduction
2. http://www.excel-easy.com/basics.html
3. http://genomicsclass.github.io/book/pages/dplyr_tutorial.html
4. http://r-statistics.co/Complete-Ggplot2-Tutorial-Part1-With-R-Code.html
5. https://www.microsoft.com/en-us/store/b/office
6. https://www.rstudio.com/products/rstudio/download/
7. https://stackoverflow.com/