ISM THE INTERNATIONAL SCHOOL OF MANAGEMENT

DBA/PhD Program

7045-SSWS - STATISTICAL SOFTWARE WORKSHOP

Final assignment: "WorldPhones"

By Laurent Dorey

- January 2018 -

Prof Teck Y. Eng

INTRODUCTION

The purpose of the workshop, and ensuing seminar was to use the R statistical programming language, to assess basic understanding of R, namely using basic commands and functions to perform data analysis and interpret basic statistical results. I have used the built-in R dataset called "World-Phones", with R-Studio for the purpose of the present assignment.

For that, I have called the "WorldPhones" datasets, trough the function "`datasets::WorldPhones`", as shared below in **Figure 1**, as built-in R Studio. Then, using a "notepad" editor, I worked with the "`write.table`[1]" function, allowing to export the "WorldPhones" datatable to a wider range of file formats instruction, e.g. CSV (or TXT, or XLS/S), exporting it from the built-in R dataset to my own folder, for better (e.g. MS-Excel, or SPSS based) management.

	N.Amer	Europe	Asia	S.Amer	Oceania	Africa	Mid.Amer
1951	45939	21574	2876	1815	1646	89	555
1956	60423	29990	4708	2568	2366	1411	733
1957	64721	32510	5230	2695	2526	1546	773
1958	68484	35218	6662	2845	2691	1663	836
1959	71799	37598	6856	3000	2868	1769	911
1960	76036	40341	8220	3145	3054	1905	1008

Figure 1: dataset "WorldPhones".

The reasons behind my choice for this Datasets was to allow for multiple row, multiple data entry combination and combined/partial extraction as I will relate below.

The very first analysis of the **Figure 1** data, is to assess the sales along the year per "zone", running the "`WorldPhones[, 1]`" instruction, so to allow to select the sole data from the column #1 (i.e. "N.Amer"), as shown on **Figure 2**, and consequently assess the standard deviation of the selected data, out of the seven years of data collection, through the instruction "`sd (World-Phones[, 1])`", that is the column #1, with the number 11.277,46 to be compared to the "mean" (see below), of 66.747,57. The Coefficient of Variation (CV = sd/mean) equaling 0,1689. It shows actually, that sales have been increasing rationally or at least gradually along the years, without years of tremendous differences. Standard deviation for "Europe" (column 2) being of 7.195,617 (mean = 34.343,43), the CV equals 0,2095 which is a bit larger than for "N. Amer", due to the faster increase between the years 1951-1956. "Asia's" (column 3) standard deviation of 2.124,215, compared to its mean of 6.229,286 provides a CV of 0,3410, showing a faster increase in sales than previously studied zones, whereas "Oceania's" (column 4) standard deviation of 496,6 876 (mean of 2.772,286) give us a CV of 0,1791 closer to the "western zones" ones. "Africa's"

[1] *write.table(WorldPhones, "C:/Laurent 2/6_DBA/7045-SSWS - Statistical Software Workshop/assignement.csv")*

(column 5), standard deviation of 523,0631 (mean of 2.625) providing with a CV of 0,19992, and "Mid.Amer" (column 6) giving us a standard deviation of 647,707 for a mean of 1.484, allowing for a CV of 0,4364 which is by far the largest one of all the seven studied. This can lead us to conclude that sales have been, in the period 1951-1966, much more dynamic (read less predictable also) in "Africa" and "Mid.Amer" than in the rest of the world, given also they low volumes.

```
 1951  1956  1957  1958  1959  1960  1961
45939 60423 64721 68484 71799 76036 79831
```

Figure 2: dataset "WorldPhones", column 1

BASIC INSTRUCTIONS

The first instruction to call in, is the one presenting all sets of data summarized, as shown in **Figure 3**. For that purpose the "`summary(WorldPhones)`" instruction is used giving us informations such as "Minimal" and "Maximal" values, per geographical zones, as well as the first and second quartiles[2] allowing for better understanding of values repartition per chunk of "25%" brackets. The "Median[3]" information allows us to know where the data sets evenly split, that is where the studied population is half below or half above the given Median value, quantity (of items) wise. It may be differing from "Average[4]" (also known as Mean) which takes into consideration the relative weight of any of the listed items. The overall amount of vehicles sold, in all the years and accross all seven zones, being equal to 805.303[5].

```
     N.Amer          Europe          Asia          S.Amer        Oceania
 Min.   :45939   Min.   :21574   Min.   :2876   Min.   :1815   Min.   :1646
 1st Qu.:62572   1st Qu.:31250   1st Qu.:4969   1st Qu.:2632   1st Qu.:2446
 Median :68484   Median :35218   Median :6662   Median :2845   Median :2691
 Mean   :66748   Mean   :34343   Mean   :6229   Mean   :2772   Mean   :2625
 3rd Qu.:73918   3rd Qu.:38970   3rd Qu.:7538   3rd Qu.:3072   3rd Qu.:2961
 Max.   :79831   Max.   :43173   Max.   :9053   Max.   :3338   Max.   :3224
                     Africa        Mid.Amer
                Min.   :  89   Min.   : 555.0
                1st Qu.:1478   1st Qu.: 753.0
                Median :1663   Median : 836.0
                Mean   :1484   Mean   : 841.7
                3rd Qu.:1837   3rd Qu.: 959.5
                Max.   :2005   Max.   :1076.0
```

Figure 3: Summary "WorldPhones" datasets

[2] *Noun: any of the three values that divide the items of a frequency distribution into four classes with each containing one fourth of the total population; also: any one of the four classes. Retrieved January 28th, 2018, from https://www.merriam-webster.com/dictionary/quartiles.*

[3] *Adjective: a value in an ordered set of values below and above which there is an equal number of values or which is the arithmetic mean of the two middle values if there is no one middle number. Retrieved January 28th, 2018, from https://www.merriam-webster.com/dictionary/median.*

[4] *Adjective: a single value (such as a mean, mode, or median) that summarizes or represents the general significance of a set of unequal values. Retrieved January 28th, 2018, from https://www.merriam-webster.com/dictionary/average.*

[5] *sum(WorldPhones[,])*

The second interesting instruction could be to add all the sales per given years on all seven zones of studies (that is "N.Amer", "Europe", "Asia", "S.Amer", "Oceania", "Africa" and "Mid.Amer"), to assess the overall yearly figures. The instruction "`rowSums(WorldPhones)`" being handy, as it gives us the information gathered on **Figure 4** across the dataset (that is per year), or the instruction "`colSums(WorldPhones)`", per geographical area, as shown on **Figure 4'.**

```
 1951   1956   1957   1958   1959   1960   1961
74494 102199 110001 118399 124801 133709 141700
```

Figure 4: Sums of "WorldPhones" datasets' rows

```
N.Amer  Europe    Asia  S.Amer Oceania  Africa Mid.Amer
467233  240404   43605   19406   18375   10388    5892
```

Figure 4': Sums of "WorldPhones" datasets' columns

GRAPHICAL INTERPRETATION & PLOTTING

While it is interesting to play with data and make a large scope of calculations, as above mentioned, it is also quite relevant to present the data in graphical ways, through their partial/total plotting. **Figure 5** below, shows the result of a set of combined instructions which allowed to present with a much more intuitive (to some extend) graphical data interpretation. Indeed, the data have been gathered per "zones" (i.e. geographical origins) and piled-up, showing accumulated sales for the 1951-1960 period of reference "`barplot(WorldPhones)`" is the first instruction to be used to build up the plot, with the use of "`space = 1`" to evenly spread the bars on the "x axe", the indication "`axes = TRUE`" to force the shaping/visual presentation of X and Y axes, "`ylab="volume""` and "`xlab="zone"`" providing the Y and X axes with their respective labels, while "`mgp=c(3.2,0.5,0)`" provides with possibilities of spacing axis label locations relative to the edge of the inner plot window. The instruction "`ylim=c(0,500000)`" was granted so to take into consideration the cumulated range of data, especially on the American market and see the bars keeping into the fame of the graph. The instructions "`legend("topright", bty = "n", legend = c("1951", "1956", "1957", "1958", "1959", "1960", "1961"), fill = c("blue", "red", "gray", "green", "yellow", "black", "white"))`" allowed for setting up of a "frameless" legend, positioned at the top right of the graph,

picking the same color code as per the bars and adding their year of collecting, while "`main= "Phone Sales"` puts an overall title on the graph.

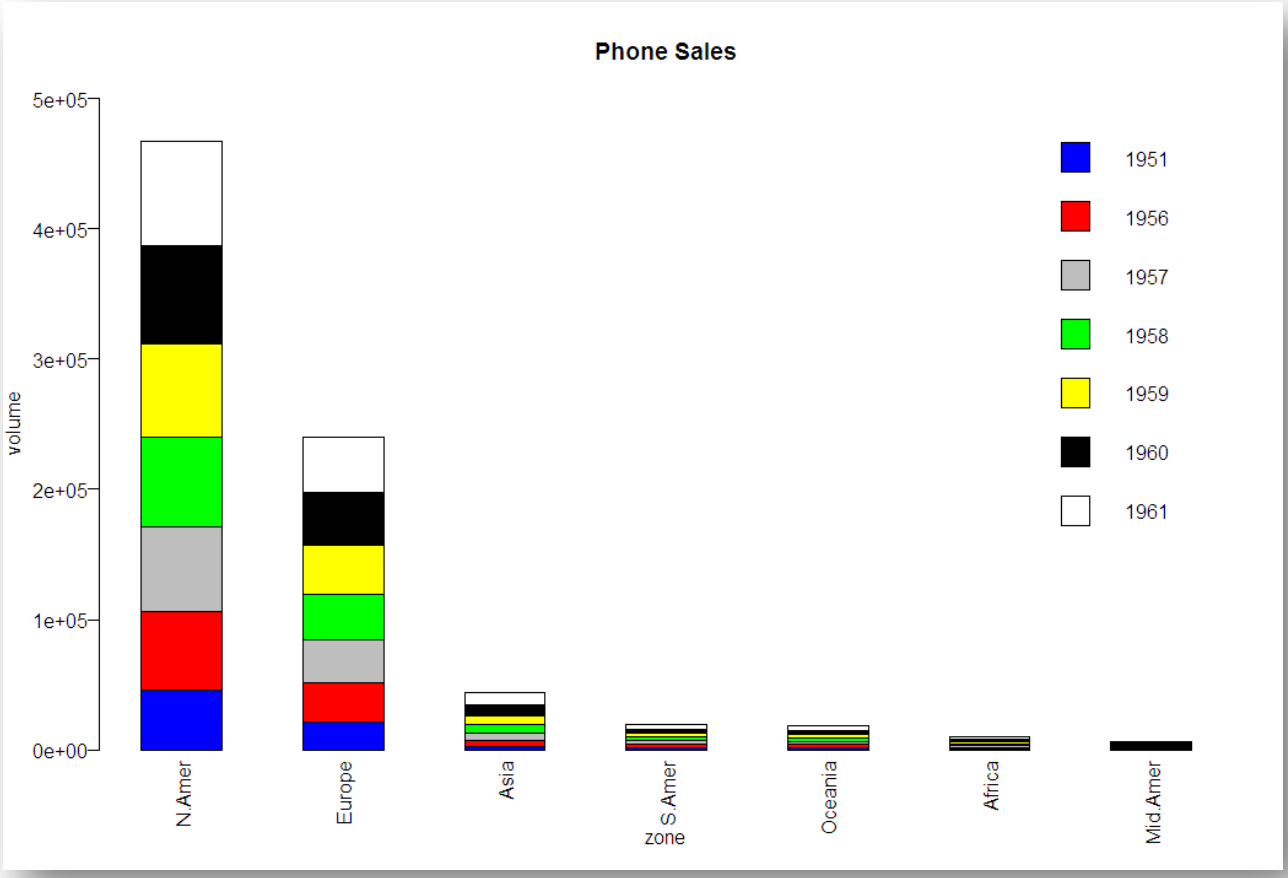

Figure 5: "WorldPhones" datasets' graphical representation

Below is the instruction set used in the making of the above presented graph.

```
> barplot(WorldPhones)
> barplot(WorldPhones,space = 1, main="Phone Sales", axes = TRUE, ylab="volume"
, las=2, xlab="zone", mgp=c(3.2,0.5,0),las=2, col= c("blue", "red", "gray", "gr
een", "yellow", "black", "white"), ylim=c(0,500000))
> legend("topright", bty = "n", legend = c("1951", "1956", "1957", "1958", "195
9", "1960", "1961"), fill = c("blue", "red", "gray", "green", "yellow", "black"
, "white"))
```

As per the previous calculation of the overall sales[6], giving 805.303 vehicles sold globally, it appeared also interesting to combine it with the Sums of "WorldPhones" datasets' columns[7], to be found on **Figure 4'** and plot it, as a pie chart, as illustrated on **Figure 6**. It allows to see the respective weight of the different geographical zones and present a coloured graph with the corresponding percentage.

[6] *see sum(WorldPhones[,])*

[7] *See "colSums(WorldPhones)"*

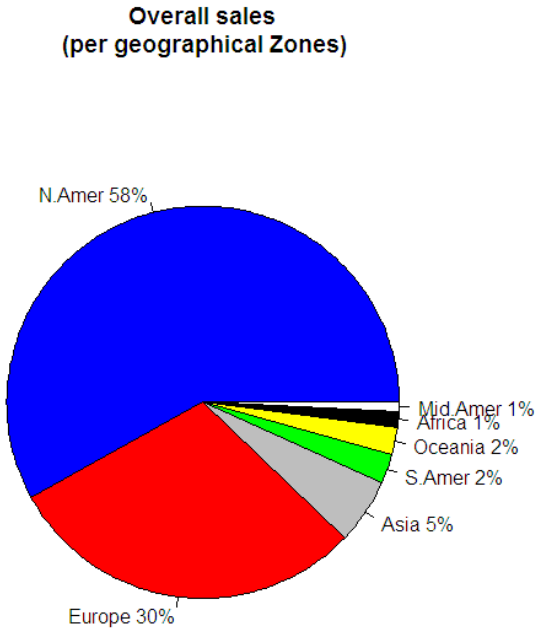

Figure 6: "WorldPhones" geographical splits

Below is the instruction set used in the making of the above presented graph.

```
> slices <- colSums(WorldPhones[,c(1,2,3,4,5,6,7)])
> lbls <- lbls <- colnames(x)
> lbls <- c("N.Amer", "Europe", "Asia", "S.Amer", "Oceania", "Africa", "Mid.Ame
r")
> pct <- round(slices/sum(slices)*100)
> lbls <- paste(lbls, pct) # add percents to labels
> lbls <- paste(lbls,"%",sep="") # ad % to labels
> pie(slices,labels = lbls, col=c("blue", "red", "gray", "green", "yellow", "bl
ack", "white"),
+     main="Overall sales\n (per geographical zones)")
```

CONCLUSION

The limited use of the R for Data Analysis and Processing application for this assignment, and the one-day Paris workshop on the topic, have shown the great potential of this "quantitative" analytical tool, yet also is relative complexity in its first-hand handling. Without a clear purpose, and thus a clear need for its use, it appeared somehow a bit difficult for me to conduct such assignment in the most comprehensive manner. However, due to the very exhaustive sets of books, blogs, online helps and other guidelines, to be found, and its very "collaborative" nature, it seems obvious that R is an application to be further assessed when time for qualitative data analysis and processing will come. "Trial and errors" were the base for such a discovery of R, yet to some extend the above mentioned plotting and basic calculation have shown information of interest, information and analysis in the need for further assessment.