Ejercicio Evaluable Tema I. Análisis Exploratorio

Bermann, M.A. & Pérez, R.S.

28 octubre, 2021

Índice

Introducción

En este informe[1] se va a proceder a desarrollar las cuestiones planteadas en el ejercicio evaluable del Tema 1 correspondiente al programa de la asignatura Técnicas Multivariantes Aplicadas al Análisis Sectorial del Máster Universitario en Modelización y Análisis de Datos Económicos (MUMADE). Para ello se va a utilizar información sobre empresas bodegueras españolas con el objetivo final de poder responder a las cuestiones mencionadas.

En un paso previo a comenzar el desarrollo de este informe es preciso definir las variables que forman parte de la base de datos de las 98 empresas bodegueras con las que vamos a trabajar.

Cuadro 1. Definición de variables

Variable	Descripción
RENECO	Rentabilidad económica (%) Últ. año disp.
RENFIN	Rentabilidad financiera (%) Últ. año disp.
LIQUIDEZ	Liquidez general (%) Últ. año disp.
ENDEUDA	Endeudamiento (%) Últ. año disp.
EMPLEA	Número de empleados. Últ. año disp.
ACTIVO	Total Activo (mil EUR) Últ. año disp.
FPIOS	Fondos propios (mil EUR) Últ. año disp.
RES	Resultado del ejercicio (mil EUR) Últ. año disp.
ING	Ingresos de explotación (mil EUR) Últ. año disp.
MARGEN	Margen de beneficio (%) Últ. año. disp.
SOLVENCIA	Coeficiente de solvencia (%) Últ. año. disp.
APALANCA	Apalancamiento (%) Últ. año disp.
FORMAJ	Forma jurídica
ACC	Número de accionistas
MATRIZ	GUO - Nombre

Los datos a utilizar en este informe, se basan en información que puede ser extraída de la base de datos Sabi, la cual contiene datos sobre empresas de España y Portugal (BVD 2021), habiéndose personalizado dichos datos en la hoja de Excel para el GRUPO_03[2] y habiendo para este informe, tal y como se ha mencionado anteriormente, un total de 98 empresas bodegueras como muestra a estudiar.

[1] Para la elaboración de este informe se ha utilizado el software R, a través de su entorno RStudio y generándose la maquetación vía R Markdown. Se han utilizado numerosas fuentes para el maquetado a partir de ayudas de Allaire et al. (2021), Cano (2021), CRAN R-Project (2021), DataCamp (2021), Keyes (2019), Luque (2019b), Luque (2019a), Van Hespen (2016), Xie, Dervieux, y Riederer (2021) y Xie, Allaire, y Grolemund (2021).

[2] GRUPO_03 es el nombre de la hoja del libro de Excel asignada para el informe.

1. Análisis de *missing values* y *outliers*

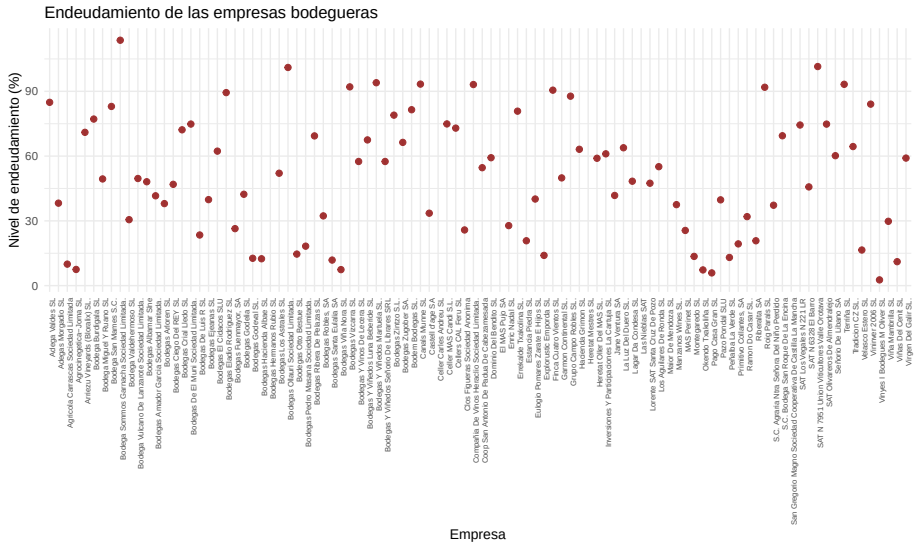
El objetivo de este primer apartado será detectar la posible existencia de *missing values* y *outliers* en la base de datos utilizada para la variable endeudamiento (`ENDEUDA`), así como decir qué casos concretos se encuentran en esta situación.

En primer lugar, podemos observar que existen 7 valores perdidos o *missing values* para el caso de la variable endeudamiento (`ENDEUDA`) (ver Cuadro 2)

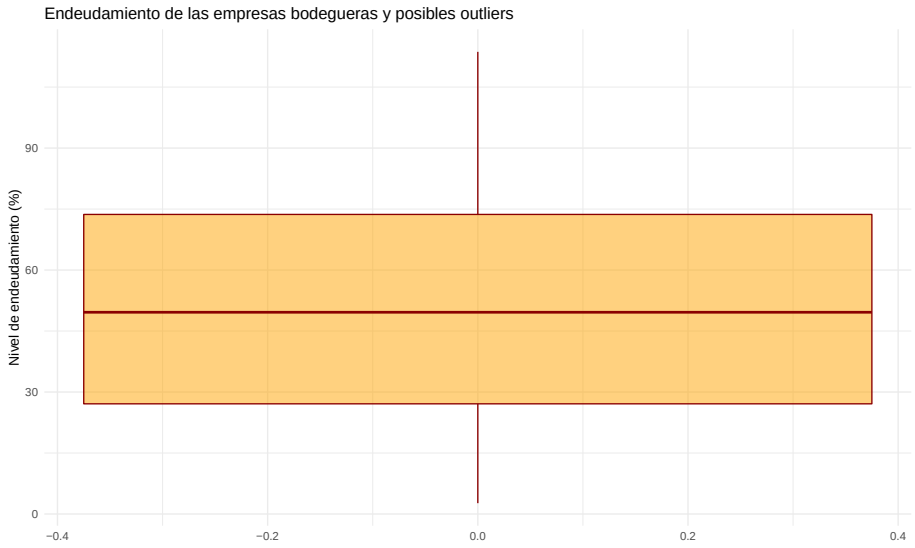
Cuadro 2: Valores perdidos en la variable de endeudamiento (ENDEUDA)

	ENDEUDA
Heretat Mestres SL	NA
Bodegas Hermanos Rubio SL	NA
San Gregorio Magno Sociedad Cooperativa De Castilla La Mancha	NA
Las Nieblas SAT	NA
Celler Carles Andreu SL	NA
Maior De Mendoza SL	NA
El MAS Pujo SA	NA

Por otro lado, el análisis de casos atípicos u *outliers* se puede realizar, en primer lugar a través de un gráfico de dispersión que nos muestra que, aparentemente, no existen dichos casos atípicos.



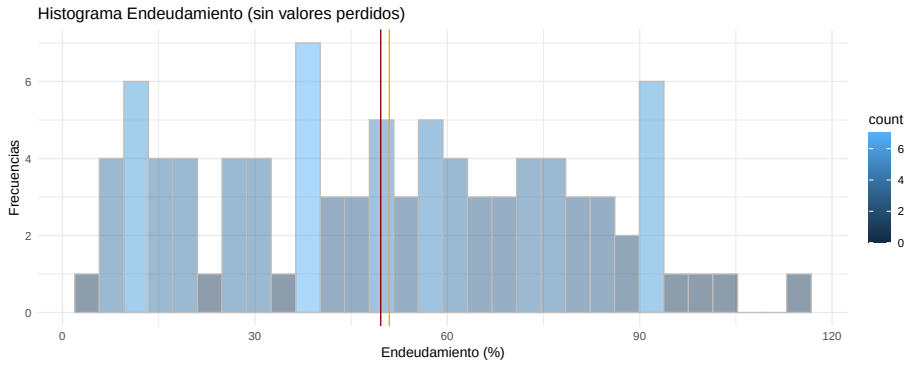
Esta posible inexistencia de casos atípicos se confirma con un gráfico `boxplot`que nos muestra, a continuación, que en la muestra no se han detectado *outliers* apreciables.


Endeudamiento de las empresas bodegueras y posibles outliers
Nivel de endeudamiento (%)
90
60
30
0
−0.4
−0.2
0.0
0.2
0.4

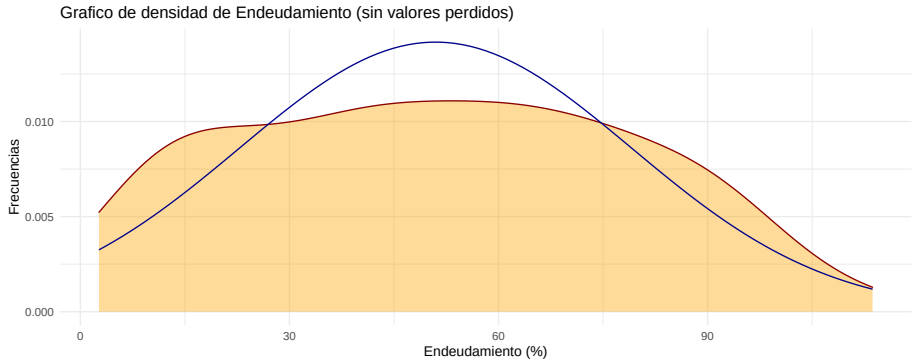
2. Caracterización de la distribución de frecuencias

En este segundo apartado procederemos a caracterizar gráficamente la distribución de frecuencias de la variable de endeudamiento (`ENDEUDA`), tanto con la base de datos manteniendo los *outliers* como eliminándolos (los *missing values* sí serán eliminados, en ambos casos).

En primer lugar, a través de un histograma, observamos la distribución de frecuencias de la variable de endeudamiento (`ENDEUDA`) de la muestra original de datos sin *missing values* (recordemos que al no existir *outliers*, tal y como se ha visto en el primer apartado, solo procede un único análisis de los datos). Dicho histograma también nos aporta información complementaria como la media y la mediana (gracias a haber eliminado los *missing values*).



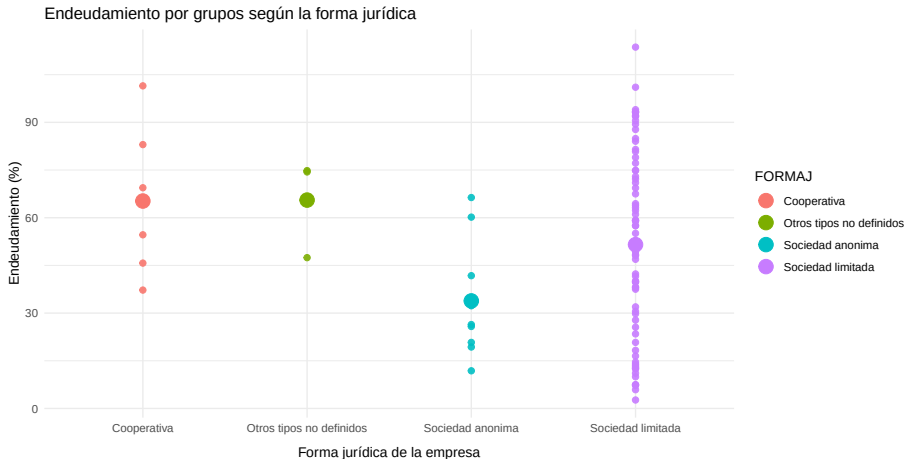
Esta información visualizada en el histograma también puede analizarse a través de un gráfico de densidad que se aporta a continuación.



3. Caracterización de la distribución de frecuencias según la forma jurídica de la empresa

En este apartado se procederá a caracterizar gráficamente (habiendo eliminado previamente los *missing values* y *outliers*) la distribución de frecuencias de la variable de endeudamiento (`ENDEUDA`) distinguiendo por la forma jurídica de las empresas (`FORMAJ`). Además se analizará si existen diferencias apreciables entre los diferentes grupos.

Para realizar el análisis se ha procedido a realizar un gráfico `geom_point`, el cual revela información de interés. Parece observarse, de forma gráfica, que las empresas bodegueras que tienen una forma jurídica basada en cooperativas, presentan niveles de endeudamiento mayores que las que se agrupan en formas jurídicas societarias (anónimas o limitadas).



4. Análisis de *missing values* y *outliers* de un conjunto de variables

El objetivo de este cuarto apartado es detectar la posible existencia de *missing values* y *outliers* en la base de datos para el caso conjunto de las variables de endeudamiento (`ENDEUDA`), ingresos de explotación (`ING`), número de empleados (`EMPLEA`) y total activo (`ACTIVO`) y señalar qué casos se encuentran en esta situación.

4.1. Análisis de *missing values*

Podemos observar que existen, concretamente, 15 filas con *missing values* para el análisis conjunto de las variables de endeudamiento (`ENDEUDA`), ingresos de explotación (`ING`), número de empleados (`EMPLEA`) y total activo (`ACTIVO`), estando todo ello reflejado en la siguiente tabla.

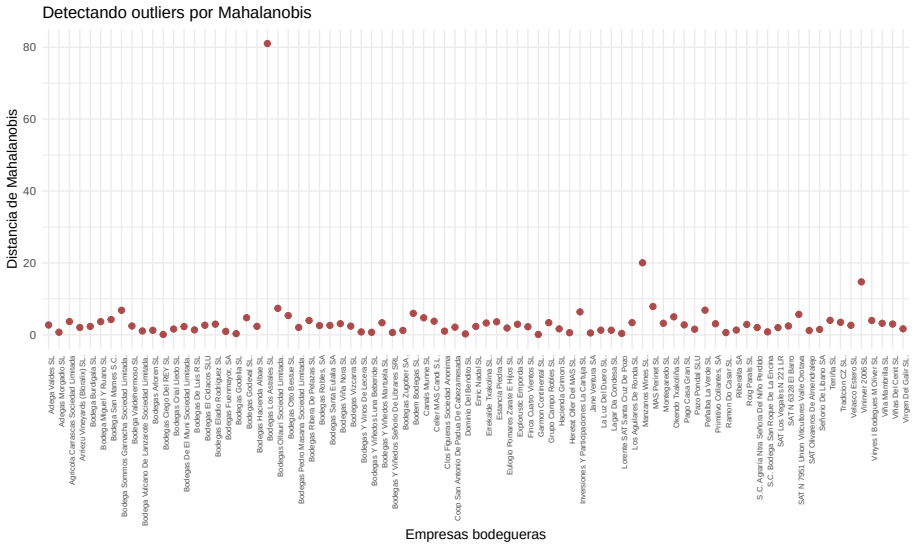
Cuadro 3: Valores perdidos en las variables de endeudamiento ('ENDEUDA'), ingresos de explotación ('ING'), número de empleados ('EMPLEA') y total activo ('ACTIVO')

	ENDEUDA	ING	EMPLEA	ACTIVO
Heretat Mestres SL	NA	870.11	NA	2484.56
Bodegas Zintzo S.L.	78.97	NA	6	NA
Bodegas Ejeanas SL	39.83	855.02	NA	NA
Bodegas Amador Garcia Sociedad Limitada.	41.61	NA	3	715.08
Bodegas Hermanos Rubio SL	NA	NA	NA	NA
Agrocinegetica-Joma SL	7.50	798.03	NA	NA
Compañia De Vinos Heraclio Sociedad Limitada.	93.15	NA	4	4789.98
San Gregorio Magno Sociedad Cooperativa De Castilla La Mancha	NA	770.35	70	865.24
Castell d'age SA	33.56	NA	3	2826.07
Las Nieblas SAT	NA	NA	4	NA
Celler Carles Andreu SL	NA	NA	NA	NA
Maior De Mendoza SL	NA	695.63	5	623.56
Cellers CAL Feru SL	72.96	695.53	NA	317.86
Bodegas Albamar Slne	48.12	694.61	2	NA
El MAS Pujo SA	NA	NA	NA	NA

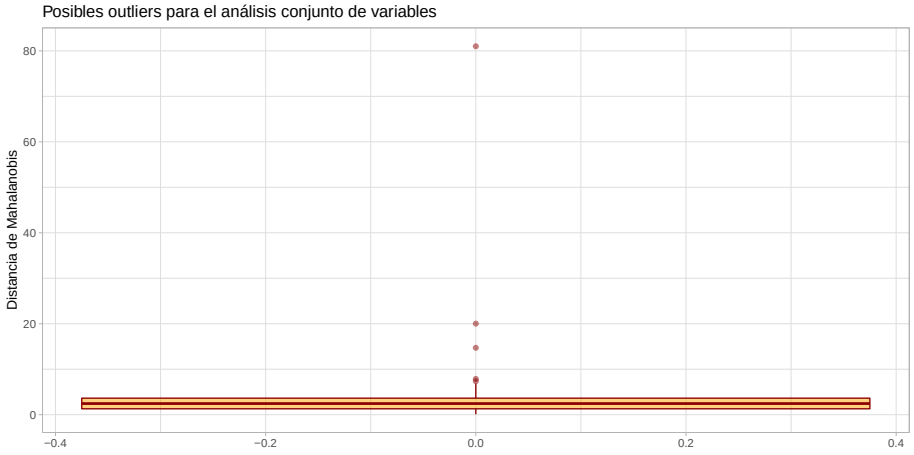
4.2. Análisis de *outliers*

Para analizar si existen *outliers* en la base de datos, para el conjunto de variables mencionadas (`ENDEUDA`, `ING`, `EMPLEA`, `ACTIVO`), se va a recurrir al análisis de las

distancias de Mahalanobis, al ser el número de variables analizadas superior a 2. Este análisis, reflejado de forma gráfica a continuación, nos muestra que habría, a primera vista, 3 casos que podrían ser claramente *outliers*.



Este análisis de las *distancias de Mahalanobis* también puede representarse a través de un gráfico *bloxplot* que, además, nos confirma que son varios los casos los que representan datos atípicos en la muestra de las empresas bodegueras para el conjunto de variables que se analiza (`ENDEUDA`, `ING`, `EMPLEA` y `ACTIVO`).



A continuación vamos a observar qué casos son los que representan datos atípicos estableciendo que cumplen esta condición aquellas empresas que tienen una

distancia de *Mahalanobis* igual o superior a 10. Así, se puede observar que estas empresas presentan ciertos datos extraordinariamente distintos respecto al resto de la muestra (ejemplo: el número de empleados de la empresa Bodegas Los Astrales SL es de 60.000, un dato muy superior al del resto de empresas).

Cuadro 4: Outliers en las variables de endeudamiento ('ENDEUDA'), ingresos de explotación ('ING'), número de empleados ('EMPLEA') y total activo ('ACTIVO')

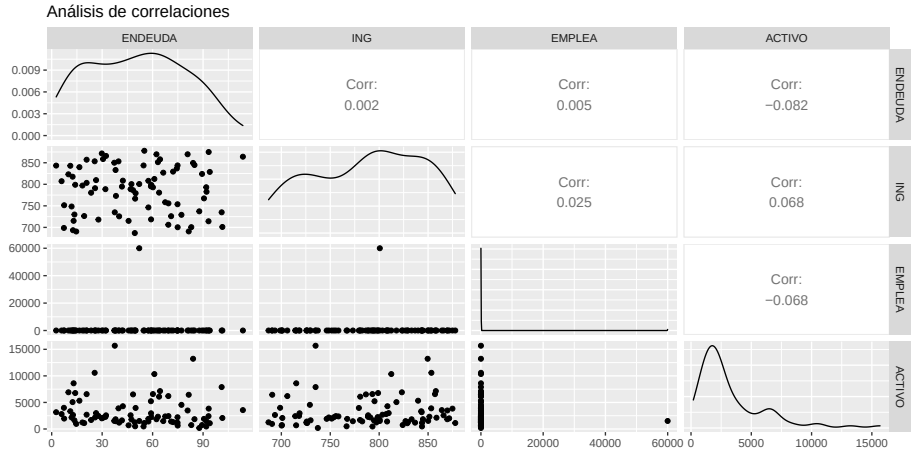
	ENDEUDA	ING	EMPLEA	ACTIVO
Vininver 2006 SL	84.067	849.5706	3	13222.471
Bodegas Los Astrales SL	52.066	800.6489	60000	1508.835
Manzanos Wines SL.	37.513	735.0001	3	15679.477

5. Análisis de correlaciones entre un conjunto de variables

En este último apartado se calculará la matriz de correlaciones entre las cuatro variables analizadas en el anterior apartado (`ENDEUDA`, `ING`, `EMPLEA` y `ACTIVO`), una vez eliminados los *missing values*, tanto en el caso de eliminar los *outliers* como en el caso de no hacerlo. También se añadirá un breve comentario sobre los resultados observados en cuanto a la relación entre las variables, así como si existen diferencias apreciables en los resultados de ambos casos.

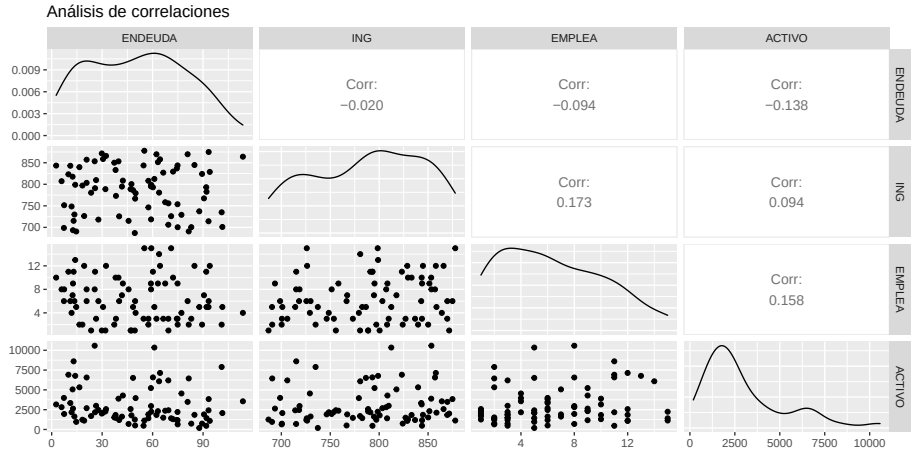
5.1. Análisis de correlaciones con *outliers*

El análisis de correlaciones de la muestra, considerando los datos también que son atípicos, nos muestra que, al existir una empresa con un extraordinariamiente elevado número de empleados, hace que los datos no sean representativos.



5.2. Análisis de correlaciones sin *outliers*

El mismo análisis, obviando los datos atípicos, nos muestran que el endeudamiento tiene poca correlación con el resto de variables, mientras que las de empleados e ingresos de explotación, y empleados y total activo es positiva pero tampoco muy elevada. En este caso, al haber eliminado los datos de las empresas que representaban *outliers*, el análisis de correlación es más coherente y cercano a la realidad.


Análisis de correlaciones
ENDEUDA
ING
EMPLEA
ACTIVO
Corr:
−0.020
Corr:
−0.094
Corr:
−0.138
Corr:
0.173
Corr:
0.094
Corr:
0.158
0.009
0.006
0.003
0.000
850
800
750
700
12
8
4
10000
7500
5000
2500
0
0
30
60
90
700
750
800
850
4
8
12
0
2500
5000
7500
10000

Referencias bibliográficas

Allaire, J. J., Rich Iannone, Alison P. Hill, y Yihui Xie. 2021. «Distill: R Markdown Format for Scientific and Technical Writing».

BVD. 2021. «Sabi». https://www.bvdinfo.com/es-es/nuestros-productos/datos/nacional/sabi.

Cano, Emilio. 2021. «Introducción al software estadístico R». https://www.lcano.com/b/iser/%7B/_%7Dbook/index.html.

CRAN R-Project. 2021. «The YAML Fieldguide». https://cran.r-project.org/web/packages/ymlthis/vignettes/yaml-fieldguide.html.

DataCamp. 2021. «RDocumentation». https://www.rdocumentation.org/.

Keyes, David. 2019. «How to make beautiful tables in R». https://rfortherestofus.com/2019/11/how-to-make-beautiful-tables-in-r/.

Luque, Pedro L. 2019a. «Cómo crear tablas de información en R Markdwon». Universidad de Sevilla. http://destio.us.es/calvo/ficheros/ComoCrearTablasRMarkdown%7B/_%7DPedroLuque%7B/_%7D2019Sep%7B/_%7Dlibrodigital.pdf.

———. 2019b. «Construcción de tablas con knitr-kableExtra».

Van Hespen, Rossana. 2016. «Writing your thesis with R Markdown (2) – Text, citations and equations». https://rosannavanhespen.nl/rmarkdown/writing-your-thesis-with-r-markdown-2-text-citations-and-equations/.

Xie, Yihui, J. J. Allaire, y Garrett Grolemund. 2021. «R Markdown: The Definitive Guide». https://bookdown.org/yihui/rmarkdown/.

Xie, Yihui, Christophe Dervieux, y Emily Riederer. 2021. «R Markdown Cookbook». https://bookdown.org/yihui/rmarkdown-cookbook/.

Anexos

Anexo 1. Código (*script*) utilizado

A continuación se presenta el *script* utilizado para desarrollar el informe

```
 [1] "---"
 [2] "title: \"Ejercicio Evaluable Tema I. AnÃ¡lisis Exploratorio\""
 [3] "author: \"Bermann, M.A. & PÃ©rez, R.S.\""
 [4] "lang: es"
 [5] "date: \"`r format(Sys.time(), '%d %B, %Y')`\""
 [6] "header-includes:"
 [7] "- \\usepackage{fancyhdr}"
 [8] "- \\pagestyle{fancy}"
 [9] "- \\fancyfoot[CO,CE]{Grupo 03 - TMAAS - MUMADE}"
[10] "- \\fancyfoot[LE,RO]{\\thepage}"
[11] "- \\usepackage{titling}"
[12] "- \\pretitle{\\begin{center}"
[13] "    \\includegraphics[width=2in,height=2in]{logo_color.png}\\LARGE\\\\}"
[14] "- \\posttitle{\\end{center}}"
[15] "documentclass: article"
[16] "bibliography: library.bib"
[17] "output:"
[18] "  pdf_document:"
[19] "    toc: yes"
[20] "---"
[21] ""
[22] "```{r setup, include=FALSE}"
[23] "knitr::opts_chunk$set(echo = TRUE, warning = FALSE, message = FALSE)"
[24] "options(width = 125)"
[25] "```"
[26] ""
[27] "```{r, echo = FALSE, include = FALSE}"
[28] "#Limpieza del entorno, activaciÃ³n de paquetes e importaciÃ³n de datos"
[29] "rm(list = ls())"
[30] "library(readxl)"
[31] "library(tidyr)"
[32] "library(knitr)"
[33] "library(flextable)"
[34] "library(magrittr)"
[35] "library(dplyr)"
[36] "library(ggplot2)"
[37] "library(GGally)"
[38] "library(kableExtra)"
[39] "bodegas_98 <- read_excel(\"tmaas_evalua_01.xlsx\", sheet = \"GRUPO_03\")"
```

```
 [40] "bodegas_98 <- data.frame(bodegas_98, row.names = 1)"
 [41] "```"
 [42] ""
 [43] "\\newpage"
 [44] ""
 [45] "# IntroducciÃ³n"
 [46] ""
 [47] "En este informe[^1] se va a proceder a desarrollar las cuestiones planteadas en el ej
 [48] ""
 [49] "En un paso previo a comenzar el desarrollo de este informe es preciso definir las var
 [50] ""
 [51] "\\begin{center}"
 [52] "Cuadro 1. DefiniciÃ³n de variables"
 [53] "\\end{center}"
 [54] "| Variable | DescripciÃ³n |"
 [55] "|----------|-------------|"
 [56] "| **RENECO** | Rentabilidad econÃ³mica (%) Ãšlt. aÃ±o disp. |"
 [57] "| **RENFIN** | Rentabilidad financiera (%) Ãšlt. aÃ±o disp. |"
 [58] "| **LIQUIDEZ**\t| Liquidez general (%) Ãšlt. aÃ±o disp. |"
 [59] "| **ENDEUDA**\t| Endeudamiento (%) Ãšlt. aÃ±o disp. |"
 [60] "| **EMPLEA**\t| NÃºmero de empleados. Ãšlt. aÃ±o disp. |"
 [61] "| **ACTIVO**\t| Total Activo (mil EUR) Ãšlt. aÃ±o disp. |"
 [62] "| **FPIOS** | Fondos propios (mil EUR) Ãšlt. aÃ±o disp. |"
 [63] "| **RES**\t| Resultado del ejercicio (mil EUR) Ãšlt. aÃ±o disp. |"
 [64] "| **ING**\t| Ingresos de explotaciÃ³n (mil EUR) Ãšlt. aÃ±o disp. |"
 [65] "| **MARGEN**\t| Margen de beneficio (%) Ãšlt. aÃ±o. disp. |"
 [66] "| **SOLVENCIA**\t| Coeficiente de solvencia (%) Ãšlt. aÃ±o. disp. |"
 [67] "| **APALANCA**\t| Apalancamiento (%) Ãšlt. aÃ±o disp. |"
 [68] "| **FORMAJ**\t| Forma jurÃdica |"
 [69] "| **ACC**\t| NÃºmero de accionistas |"
 [70] "| **MATRIZ** | GUO - Nombre |"
 [71] ""
 [72] "Los datos a utilizar en este informe, se basan en informaciÃ³n que puede ser extraÃ-
da de la base de datos Sabi, la cual contiene datos sobre empresas de EspaÃ±a y Portugal [@B
 [73] ""
 [74] "[^1]: Para la elaboraciÃ³n de este informe se ha utilizado el software R, a travÃ©s d
a R Markdown. Se han utilizado numerosas fuentes para el maquetado a partir de ayudas de @Al
 [75] "[^2]: GRUPO_03 es el nombre de la hoja del libro de Excel asignada para el informe."
 [76] ""
 [77] "# 1. AnÃ¡lisis de _missing values_ y _outliers_"
 [78] ""
 [79] "El objetivo de este primer apartado serÃ¡ detectar la posible existencia de _missing
como decir quÃ© casos concretos se encuentran en esta situaciÃ³n."
 [80] ""
 [81] "En primer lugar, podemos observar que existen 7 valores perdidos o _missing values_ p
 [82] "```{r, echo = FALSE}"
```

```
 [83] "#detectando missing values"
 [84] "bodegas_98 %>% "
 [85] "  filter(is.na(ENDEUDA)) %>% "
 [86] "  select(ENDEUDA) %>% "
 [87] "  kable(caption = \"Valores perdidos en la variable de endeudamiento (ENDEUDA)\") %>%
 [88] "  kable_styling(font_size = 8,"
 [89] "                latex_options = c(\"striped\", \"HOLD_position\"), "
 [90] "                full_width = T, position = \"center\")"
 [91] "‘‘‘‘"
 [92] ""
 [93] "Por otro lado, el anÃ¡lisis de casos atÃpicos u _outliers_ se puede realizar, en prin
picos."
 [94] ""
 [95] "‘‘‘‘{r, echo = FALSE, fig.align = 'center', fig.width = 10, fig.height = 6}"
 [96] "#anÃ¡lisis grÃ¡fico de outliers 1: geom_point para detectar outliers"
 [97] "ggplot(data = bodegas_98, aes(x = row.names(bodegas_98), y = ENDEUDA)) +"
 [98] "  geom_point(size = 2, alpha = 0.8, colour = 'red4') +"
 [99] "  xlab('Empresa') +"
[100] "  ylab('Nivel de endeudamiento (%)') +"
[101] "  ggtitle('Endeudamiento de las empresas bodegueras') +"
[102] "  theme_minimal() +"
[103] "  theme(axis.text.x = element_text(angle = 90, size = 6, hjust = 1, vjust = 1))"
[104] "‘‘‘‘"
[105] ""
[106] "Esta posible inexistencia de casos atÃpicos se confirma con un grÃ¡fico ‘boxplot‘que
[107] ""
[108] "‘‘‘‘{r, echo = FALSE, fig.align = 'center', fig.width = 10, fig.height = 6}"
[109] "#anÃ¡lisis grÃ¡fico de outliers 2: bloxplot para detectar outliers"
[110] "ggplot(data = bodegas_98, aes(y = ENDEUDA)) +"
[111] "  geom_boxplot(alpha = 0.5, fill = \"orange\", color = \"red4\") +"
[112] "  ylab('Nivel de endeudamiento (%)') +"
[113] "  ggtitle('Endeudamiento de las empresas bodegueras y posibles outliers') +"
[114] "  theme_minimal()"
[115] "‘‘‘‘"
[116] ""
[117] "\\newpage"
[118] ""
[119] "# 2. CaracterizaciÃ³n de la distribuciÃ³n de frecuencias"
[120] ""
[121] "En este segundo apartado procederemos a caracterizar grÃ¡ficamente la distribuciÃ³n d
serÃ¡n eliminados, en ambos casos)."
[122] ""
[123] "‘‘‘‘{r, echo = FALSE}"
[124] "#creando un nuevo data.frame para poder conservar el original"
[125] "bodegas_muestra1 <- select(bodegas_98, everything())"
[126] ""
```

```
[127] "#eliminando los missing values en el nueva data.frame"
[128] "bodegas_muestra1 <- bodegas_muestra1 %>% "
[129] "  filter(! is.na(ENDEUDA))"
[130] "‘‘‘"
[131] ""
[132] "En primer lugar, a travÃ©s de un histograma, observamos la distribuciÃ³n de frecuenci
[133] ""
[134] "‘‘‘{r, echo = FALSE, fig.align = 'center', fig.width = 10, fig.height = 4}"
[135] "ggplot(data = bodegas_muestra1, aes(x = ENDEUDA)) +"
[136] "  geom_histogram(color='grey', aes(fill=..count..), alpha = 0.5) +"
[137] "  geom_vline(xintercept = mean(bodegas_muestra1$ENDEUDA), color = \"orange\") +"
[138] "  geom_vline(xintercept = median(bodegas_muestra1$ENDEUDA), color = \"red4\") +"
[139] "  xlab('Endeudamiento (%)') +"
[140] "  ylab('Frecuencias') +"
[141] "  ggtitle('Histograma Endeudamiento (sin valores perdidos)') +"
[142] "  theme_minimal()"
[143] "‘‘‘"
[144] ""
[145] "Esta informaciÃ³n visualizada en el histograma tambiÃ©n puede analizarse a travÃ©s de
[146] ""
[147] "‘‘‘{r, echo = FALSE, fig.align = 'center', fig.width = 10, fig.height = 4}"
[148] "ggplot(data= bodegas_muestra1, aes(x = ENDEUDA)) +"
[149] "  geom_density(alpha = 0.4, fill = \"orange\", color = \"red4\") +"
[150] "  xlab('Endeudamiento (%)') +"
[151] "  ylab('Frecuencias') +"
[152] "  ggtitle('Grafico de densidad de Endeudamiento (sin valores perdidos)') +"
[153] "  stat_function(fun = dnorm, color=\"blue4\","
[154] "                args = list(mean = mean(bodegas_muestra1$ENDEUDA),"
[155] "                            sd = sd(bodegas_muestra1$ENDEUDA))) +"
[156] "  theme_minimal()"
[157] "‘‘‘"
[158] ""
[159] "\\newpage"
[160] ""
[161] "# 3. CaracterizaciÃ³n de la distribuciÃ³n de frecuencias segÃºn la forma jurÃ-
dica de la empresa"
[162] ""
[163] "En este apartado se procederÃ¡ a caracterizar grÃ¡ficamente (habiendo eliminado previ
dica de las empresas (‘FORMAJ‘). AdemÃ¡s se analizarÃ¡ si existen diferencias apreciables en
[164] ""
[165] "Para realizar el anÃ¡lisis se ha procedido a realizar un grÃ¡fico ‘geom_point‘, el cu
dica basada en cooperativas, presentan niveles de endeudamiento mayores que las que se agrup
dicas societarias (anÃ³nimas o limitadas)."
[166] ""
[167] "‘‘‘{r, echo = FALSE, fig.align = 'center', fig.width = 10, fig.height = 5}"
[168] "ggplot(data = bodegas_muestra1, aes(x=FORMAJ, y = ENDEUDA)) +"
```

```
[169] "  geom_point(aes (group = FORMAJ, color = FORMAJ), size = 2, alpha = 0.9) +"
[170] "  xlab('Forma jurÃdica de la empresa') +"
[171] "  ylab('Endeudamiento (%)') +"
[172] "  ggtitle('Endeudamiento por grupos segÃºn la forma jurÃ-
dica') +"
[173] "  stat_summary(fun = \"mean\", geom = \"point\", size = 5, aes(col = FORMAJ)) +"
[174] "  theme_minimal() +"
[175] "  theme(axis.title.x = element_text(vjust = -2))"
[176] "````"
[177] ""
[178] "\\newpage"
[179] ""
[180] "# 4. AnÃ¡lisis de _missing values_ y _outliers_ de un conjunto de variables"
[181] ""
[182] "El objetivo de este cuarto apartado es detectar la posible existencia de _missing val
[183] ""
[184] "````{r, echo = FALSE, include = FALSE}"
[185] "#creando un nuevo data.frame para poder conservar el original"
[186] "bodegas_muestra2 <- select(bodegas_98, everything())"
[187] "````"
[188] ""
[189] "## 4.1. AnÃ¡lisis de _missing values_"
[190] ""
[191] "Podemos observar que existen, concretamente, 15 filas con _missing values_ para el an
[192] ""
[193] "````{r, echo = FALSE}"
[194] "# detectando missing values"
[195] "bodegas_muestra2 %>%"
[196] "  filter(is.na(ENDEUDA) | is.na(ING) | is.na(EMPLEA) | is.na(ACTIVO)) %>%"
[197] "  select(ENDEUDA, ING, EMPLEA, ACTIVO) %>%"
[198] "  kable(caption = \"Valores perdidos en las variables de endeudamiento (`ENDEUDA`), i
[199] "  kable_styling(font_size = 5,"
[200] "                latex_options = c(\"striped\", \"HOLD_position\", \"scale_down\"), "
[201] "                full_width = T, position = \"center\")"
[202] "````"
[203] ""
[204] "````{r, echo = FALSE, include = FALSE}"
[205] "#eliminando missing values en la muestra"
[206] "bodegas_muestra2 <- bodegas_muestra2 %>% "
[207] "  filter(! is.na(ENDEUDA) & ! is.na(ING) & ! is.na(EMPLEA) & ! is.na(ACTIVO))"
[208] "````"
[209] ""
[210] "## 4.2. AnÃ¡lisis de _outliers_"
[211] ""
[212] "Para analizar si existen _outliers_ en la base de datos, para el conjunto de variable
a, a primera vista, 3 casos que podrÃan ser claramente _outliers_."
```

```
[213] ""
[214] "‘‘‘{r, echo = FALSE, fig.align = ’center’, fig.width = 10, fig.height = 6}"
[215] "#detectando outliers analizando el vector de las variables"
[216] "bodegas_muestra2_maha <- bodegas_muestra2 %>%"
[217] "  select(ENDEUDA, ING, EMPLEA, ACTIVO)"
[218] "maha_bodegas <- mahalanobis(bodegas_muestra2_maha[,1:4],"
[219] "                            center = colMeans(bodegas_muestra2_maha[,1:4]),"
[220] "                            cov = cov(bodegas_muestra2_maha[,1:4]))"
[221] ""
[222] "# grÃ¡fico distancia de mahalanobis con geom_point"
[223] "ggplot(data = bodegas_muestra2, aes(x = row.names(bodegas_muestra2), y = maha_bodegas
[224] "  geom_point(size = 2, alpha = 0.7, color=’red4’) +"
[225] "  xlab(’Empresas bodegueras’) +"
[226] "  ylab(’Distancia de Mahalanobis’) +"
[227] "  ggtitle(’Detectando outliers por Mahalanobis’) +"
[228] "  theme_minimal() +"
[229] "  theme(axis.text.x = element_text(angle = 90, size = 6,hjust = 1, vjust = 1))"
[230] "‘‘‘‘"
[231] ""
[232] "Este anÃ¡lisis de las _distancias de Mahalanobis_ tambiÃ©n puede representarse a trav
picos en la muestra de las empresas bodegueras para el conjunto de variables que se analiza
[233] ""
[234] "‘‘‘{r, echo = FALSE, fig.align = ’center’, fig.width = 10, fig.height = 5}"
[235] "# grÃ¡fico distancias de mahalanobis con bloxplot"
[236] "ggplot(data = bodegas_muestra2, aes(y = maha_bodegas)) +"
[237] "  geom_boxplot(alpha = 0.5, fill = \"orange\", color = \"red4\") +"
[238] "  ylab(’Distancia de Mahalanobis’) +"
[239] "  ggtitle(’Posibles outliers para el anÃ¡lisis conjunto de variables’) +"
[240] "  theme_light()"
[241] "‘‘‘‘"
[242] ""
[243] "A continuaciÃ³n vamos a observar quÃ© casos son los que representan datos atÃ-
picos estableciendo que cumplen esta condiciÃ³n aquellas empresas que tienen una distancia c
, se puede observar que estas empresas presentan ciertos datos extraordinariamente distintos
[244] ""
[245] "‘‘‘{r, echo = FALSE}"
[246] "# eliminando outliers"
[247] "bodegas_muestra2 %>% filter(maha_bodegas >= 10) %>%"
[248] "  select(ENDEUDA, ING, EMPLEA, ACTIVO) %>%"
[249] "  kable(caption = \"Outliers en las variables de endeudamiento (‘ENDEUDA‘), ingresos
[250] "  kable_styling(font_size = 8,"
[251] "                latex_options = c(\"striped\", \"HOLD_position\"), "
[252] "                full_width = T, position = \"center\")"
[253] ""
[254] "#nuevo vector sin los outliers"
[255] "bodegas_muestra3 <- bodegas_muestra2 %>%"
```

```
[256] "  filter(maha_bodegas < 10) "
[257] "‘‘‘‘"
[258] ""
[259] "\\newpage"
[260] ""
[261] "# 5. AnÃ¡lisis de correlaciones entre un conjunto de variables"
[262] ""
[263] "En este Ãºltimo apartado se calcularÃ¡ la matriz de correlaciones entre las cuatro va
como si existen diferencias apreciables en los resultados de ambos casos."
[264] ""
[265] "## 5.1. AnÃ¡lisis de correlaciones con _outliers_"
[266] ""
[267] "El anÃ¡lisis de correlaciones de la muestra, considerando los datos tambiÃ©n que son
picos, nos muestra que, al existir una empresa con un extraordinariamiente elevado nÃºmero d
[268] ""
[269] "‘‘‘‘{r, echo = FALSE, fig.align = ’center’, fig.width = 10, fig.height = 5}"
[270] "#no eliminando outliers"
[271] "bodegas_muestra2_cor <- bodegas_muestra2 %>%"
[272] "  select(ENDEUDA, ING, EMPLEA, ACTIVO)"
[273] "ggpairs(bodegas_muestra2_cor, title = \"AnÃ¡lisis de correlaciones\")"
[274] "‘‘‘‘"
[275] ""
[276] "## 5.2. AnÃ¡lisis de correlaciones sin _outliers_"
[277] ""
[278] "El mismo anÃ¡lisis, obviando los datos atÃpicos, nos muestran que el endeudamiento ti
[279] ""
[280] "‘‘‘‘{r, echo = FALSE, fig.align = ’center’, fig.width = 10, fig.height = 5}"
[281] "#correlaciones con eliminando outliers"
[282] "bodegas_muestra3_cor <- bodegas_muestra3 %>%"
[283] "  select(ENDEUDA, ING, EMPLEA, ACTIVO)"
[284] "ggpairs(bodegas_muestra3_cor, title = \"AnÃ¡lisis de correlaciones\")"
[285] "‘‘‘‘"
[286] ""
[287] "\\newpage"
[288] ""
[289] "# Referencias bibliogrÃ¡ficas"
[290] ""
[291] "<div id=\"refs\"></div>"
[292] ""
[293] "\\newpage"
[294] ""
[295] "# Anexos"
[296] ""
[297] "## Anexo 1. CÃ³digo (_script_) utilizado"
[298] ""
[299] "A continuaciÃ³n se presenta el _script_ utilizado para desarrollar el informe"
```

```
[300] ""
[301] "‘‘‘{r, echo = FALSE, comment= ’’}"
[302] "script <- readLines(\"TMAAS_01.Rmd\")"
[303] "print(script)"
[304] "‘‘‘‘"
[305] ""
[306] "\\newpage"
[307] ""
[308] "## Anexo 2. Datos de la sesiÃ³n"
[309] ""
[310] "En esta secciÃ³n se recogen los datos de la sesiÃ³n utilizada para elaborar este info
como las versiones de los paquetes bajo los cuales se ha ejecutado el cÃ³digo o _script_."
[311] ""
[312] "‘‘‘{r, echo = FALSE, comment = ’’}"
[313] "sessionInfo()"
[314] "‘‘‘‘"
```

Anexo 2. Datos de la sesión

En esta sección se recogen los datos de la sesión utilizada para elaborar este informe. Siguiendo a Cano (2021), es fundamental observar la versión de R, así como las versiones de los paquetes bajo los cuales se ha ejecutado el código o *script*.

```
R version 4.1.1 (2021-08-10)
Platform: x86_64-w64-mingw32/x64 (64-bit)
Running under: Windows 10 x64 (build 19042)

Matrix products: default

locale:
[1] LC_COLLATE=Spanish_Spain.1252  LC_CTYPE=Spanish_Spain.1252    LC_MONETARY=Spanish_Spain.
[4] LC_NUMERIC=C                   LC_TIME=Spanish_Spain.1252

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
[1] kableExtra_1.3.4 GGally_2.1.2     ggplot2_3.3.5    dplyr_1.0.7      magrittr_2.0.1   fle
[8] tidyr_1.1.4      readxl_1.3.1

loaded via a namespace (and not attached):
 [1] tidyselect_1.1.1   xfun_0.26          purrr_0.3.4        colorspace_2.0-2   vctrs_0.3.8
 [7] viridisLite_0.4.0  htmltools_0.5.2    yaml_2.2.1         base64enc_0.1-3    utf8_1.2.2
[13] pillar_1.6.3       glue_1.4.2         withr_2.4.2        DBI_1.1.1          gdtools_0.2
[19] uuid_0.1-4         lifecycle_1.0.1    plyr_1.8.6         stringr_1.4.0      munsell_0.5
[25] cellranger_1.1.0   rvest_1.0.1        zip_2.2.0          evaluate_0.14      labeling_0.
[31] fansi_0.5.0        Rcpp_1.0.7         scales_1.1.1       webshot_0.5.2      farver_2.1.
[37] digest_0.6.28      stringi_1.7.4      grid_4.1.1         tools_4.1.1        tibble_3.1.
[43] pkgconfig_2.0.3    ellipsis_0.3.2     data.table_1.14.2  xml2_1.3.2         svglite_2.0
[49] assertthat_0.2.1   rmarkdown_2.11     reshape_0.8.8      officer_0.4.0      rstudioapi_
[55] compiler_4.1.1
```