REPRODUCIBILIDAD BASADO EN R

REPORTE

3R Project

1 Introducción

El presente informe forma parte del Semillero de Investigación 3R Project: Reproducibilidad basada en R. Este proyecto tiene como objetivo principal transformar investigaciones científicas en documentos reproducibles, promoviendo la transparencia y la verificabilidad en el ámbito científico.

En este documento se detalla el proceso de exploración y análisis de un conjunto de datos, utilizando herramientas de R para garantizar la reproducibilidad de cada etapa

Durante la revisión de la data y su comparación con el informe original, se identificaron diversas inconsistencias que requieren un análisis detallado. Una de las primeras observaciones es que las variables en la base de datos no están en el mismo idioma que las presentadas en el informe, lo que puede generar dificultades en la interpretación y procesamiento de la información.

Asimismo, se detectó una discrepancia en el número total de observaciones. Según la Tabla 2 del informe, el total de registros reportados es menor en comparación con los datos almacenados en la base de datos (Excel). En concreto, la diferencia total es de 134 observaciones menos en el informe, con una reducción específica de 74 registros en España y 60 en Italia.

Otro aspecto crítico identificado es la presencia de registros con valores faltantes o inválidos en la base de datos, tanto en Excel como en R. En Excel, se presentan errores como N/A para casillas vacías de tipo numérico, #NUM! cuando se intenta calcular logaritmos de números negativos, casillas en blanco para valores de tipo texto y #¡DIV/0! cuando se realizan operaciones con celdas vacías. En R, los errores más comunes incluyen NAN, que indica casillas vacías, y Inf, que aparece cuando se intenta realizar un cálculo con valores nulos o se produce un error en la operación.

Estas observaciones iniciales evidencian la necesidad de llevar a cabo un proceso de limpieza, estandarización y verificación de la data antes de continuar con el análisis. Garantizar la coherencia y fiabilidad de los datos es fundamental para obtener resultados precisos y reproducibles.

2 Instalación de Paquetes

Estos paquetes se instalaran para llevar a cabo un flujo de trabajo reproducible en la exploración de datos. tidyverse permite manipular, limpiar y visualizar los datos; readr facilita su importación eficiente; y janitor ayuda a organizarlos y limpiarlos. Para documentar y presentar los resultados, rstatix realiza análisis estadísticos, mientras que gt genera tablas profesionales y openxlsx exporta los reportes a Excel. En conjunto, estos paquetes garantizan un análisis reproducible, claro y bien documentado.

```
install.packages("tidyverse")
install.packages("openxlsx")
install.packages("readr")
install.packages("janitor")
```

```r
install.packages("rstatix")
install.packages("gt")
```

3 Carga de Librerias

Estas líneas de código en R sirven para cargar paquetes previamente instalados para utilizarlos en el análisis. Cada library() activa el paquete especificado, permitiendo el acceso a sus funciones. Aquí te explico brevemente qué hace cada uno:

```r
library(tidyverse)
library(dplyr)
library(openxlsx)
library(readr)
library(janitor)
library(rstatix)
library(gt)
```

4 Funciones Especificas

Estas son funciones específicas del paquete base stats en R, y están relacionadas con análisis estadísticos. El prefijo stats:: se usa para asegurarse de que se está llamando directamente a las funciones del paquete stats, incluso si hay otras funciones con el mismo nombre en otros paquetes cargados

```r
stats::chisq.test
stats::fisher.test
stats::filter
```

5 Importando Datos

En esta sección, se carga el conjunto de datos en un objeto de R utilizando la función read.xlsx. Esto permite explorar y manipular los datos para análisis posteriores. La finalidad es tener acceso al archivo original y comenzar con la limpieza y transformación de los datos

```r
library(openxlsx)
datos<-read.xlsx("C:/Users/Usuario/Desktop/SEMILLERO/DATA_STATA/DATA/DatosStata.xlsx")
```

6 Explorando el Objeto de los Datos

Aquí se realiza una inspección básica del objeto de datos con la función str para entender su estructura y tipos de variables.

```r
str(datos)
```

7 Eliminando Duplicados

También se seleccionan y eliminan columnas irrelevantes con select del paquete dplyr, optimizando el dataset para el análisis. Esto asegura que solo se trabajen datos útiles para los objetivos del proyecto.

```
datos_ex <-datos %>% select (- c(Número.de.accionistas,DM.Edad,ADV.Edad))
```

8 Diccionario de Traducción

En esta fase estudio variables del conjunto de datos de la Data que se encontraba en español. Dado que los datos originales estaban en español, se identificaron las etiquetas y nombres de las variables que necesitaban ser traducidas para alinearlas con las utilizadas en el artículo original, que estaba redactado en inglés. Lo que representaba un desafío para la comprensión de las variables y su correspondencia con las mencionadas en el documento fuente. Para garantizar una mayor claridad y alineación conceptual, se procedió a traducir las etiquetas y nombres de las variables del archivo Excel del español al inglés. Este paso fue esencial para comprender mejor las características y contextos de las variables del artículo y su relación con los datos entregados.

La traducción no solo facilitó la interpretación de los datos, sino que también permitió identificar posibles discrepancias o inconsistencias entre la documentación original y los datos proporcionados. Esta etapa inicial fue fundamental para establecer las bases de un análisis robusto y coherente con los objetivos del artículo académico."

```
DATA <- datos_ex %>%
  rename(
    "Year" = "Año",
    "ID" = "Id.",
    "Identity" = "Ident",
    "Gender" = "Género",
    "Company_Name" = "Nombre.empresa",
    "City" = "Ciudad",
    "Country_ISO_Code" = "Código.ISO.del.país",
    "Total_Assets" = "Activos.totales",
    "Growth" = "Grow",
    "GDP" = "PIB",
    "GDP_Var" = "VarPIB",
    "Inflation" = "Inflacion",
    "Ln_GDP" = "LnPIB",
    "Ln_Inflation" = "LnInflacion",
    "Country" = "Pais",
    "NACE_Code" = "NACE.code",
    "Employees_Last_Year" = "Número.empleados.Últ..año.disp.",
    "Standard_Legal_Form" = "Forma.jurídica.estándar",
    "Legal_Form_Tabul" = "FJurídicaTabul",
    "Legal_Form" = "FormaJurídica",
    "Incorporation_Date" = "Fecha.de.constitución",
    "End_Date" = "Fecha.final",
    "Seniority" = "Antigüedad",
    "Ln_Seniority" = "LnAntigüedad",
    "Cash_Flows" = "Flujos.de.Caja",
    "Fixed_Assets" = "Activos.Fijos",
    "Current_Assets" = "Activos.Corrientes",
```

```
    "Inventory" = "Stock",
    "Receivables" = "Deudores",
    "Other_Current_Assets" = "Otros.activos.corrientes",
    "Cash_and_Equivalents" = "Efectivo.y.equivalentes",
    "Ln_Fixed_Assets" = "LnActFijo",
    "Ln_Current_Assets" = "LnActCorr",
    "Ln_Inventory" = "LnStock",
    "Ln_Receivables" = "LnDeudores",
    "Ln_Other_Assets" = "LnOtrosactiv",
    "Ln_Cash" = "LnEfectivo",
    "Ln_Total_Assets" = "LnActTotal",
    "Non_Current_Liabilities" = "Pasivos.no.corrientes",
    "Current_Liabilities" = "Pasivos.Corrientes",
    "Liquidity1" = "Liquidez1",
    "Liquidity1_Dummy" = "Liquidez1Dummy",
    "Total_Liabilities" = "Pasivo.Total",
    "Equity" = "Fondos.Propios",
    "Ln_Non_Current_Liabilities" = "LnPasivoNoCorr",
    "Ln_Current_Liabilities" = "LnPasivoCorr",
    "Ln_Total_Liabilities" = "LnPasivoTotal",
    "Ln_Equity" = "LnFondosPropios",
    "Operating_Revenue" = "Ingresos.Explotación",
    "Operating_Profit" = "Resultado.Explotación",
    "Financial_Expenses" = "Gastos.Financieros",
    "Ordinary_Profit_Before_Tax" = "Rdo..Ordinario.antes.Impuestos",
    "Taxes" = "Impuestos",
    "Ordinary_Activities_Profit" = "Rdo..Actividades.Odinarias",
    "Extraordinary_and_Other_Profit" = "Rdo..Extr..y.Otros",
    "Net_Profit" = "Rdo.Ejercicio",
    "ROE" = "ROE",
    "ROA" = "ROA",
    "Collection_Period" = "Período.de.Cobro",
    "Credit_Period" = "Período.de.Credito",
    "ROEE" = "ROEE",
    "ROAA" = "ROAA",
    "CollectionPeriod" = "PeríodoCobro",
    "Payment_Period" = "PeríodoPago",
    "PMC_PMP" = "PMC-PMP",
    "Net_Asset_Turnover" = "Rotación.de.activos.netos",
    "Inventory_Turnover" = "Rotación.de.las.existencias",
    "Solvency_Turnover" = "Rotacion.de.Solvencia",
    "Asset_Turnover" = "RotacActivos",
    "InventoryTurnover" = "RotacExistenc",
    "SolvencyTurnover" = "RotacSolvencia",
    "Liquidity_Ratio" = "Ratio.de.Liquidez",
    "Leverage" = "Apalancamiento",
    "Profit_per_Employee" = "Beneficio.por.empleado",
    "Operating_Revenue_per_Employee" = "Ingresos.Explotación.por.empleado",
    "Average_Employee_Cost" = "Coste.medio.Empleados",
    "Total_Assets_per_Employee" = "Total.acivos.por.empleado",
    "Levera" = "Apalancam",
    "Profit_Employee" = "Benefic/empleado",
    "OperatingRevenue_Employee" = "IngrExpl/empleado",
```

```
    "Cost_Employee" = "Coste/empleado",
    "Assets_Employee" = "Activos/empleado",
    "Number_of_Board_and_Management_Members" =
      "Nümero.de.miembros.de.las.juntas.&.gestión",
    "Board_Members" = "MiembrosJuntas",
    "DM_Full_Name" = "DM.Nombre.completo",
    "DM_Job_Title" = "DM.Título.trabajo",
    "Shareholder_Direct_Percentage" = "Accionista.-.%.directo",
    "Shareholder_Total_Percentage" = "Accionista.-.%.total",
    "CSH_Direct_Percentage" = "CSH.-.%.directo",
    "DM_Original_Job_Title" = "DM.Título.original.trabajo",
    "DM_Board_Committee_or_Executive_Department" =
      "DM.Junta,.comité.or.departamento.ejecutivo",
    "DM_Level_of_Responsibility" = "DM.Nivel.de.responsabilidad",
    "DM_First_Name" = "DM.Nombre",
    "DM_Last_Name" = "DM.Apellido",
    "DM_Gender" = "DM.Género",
    "DM_Nationality_Country" = "DM.País.de.nacionalidad",
    "DM_Also_a_Shareholder" = "DM.También.un.accionista",
    "DM_Position_Type" = "DM.Tipo.de.posición",
    "Number_of_Advisors" = "Número.de.asesores",
    "ADV_First_Name" = "ADV.Nombre",
    "ADV_Last_Name" = "ADV.Apellido",
    "ADV_Gender" = "ADV.Género",
    "ADV_Nationality_Country" = "ADV.País.de.nacionalidad",
    "Nationality_Country" = "País.de.nacionalidad",
    "Number_of_Employees" = "Número.empleados",
    "BvD_Independence_Indicator" = "Indicador.independencia.BvD"
  )
```

9 Coercion de datos

Las funciones de coerción en R son fundamentales para transformar los tipos de datos y garantizar que sean compatibles con los análisis o manipulaciones que se desean realizar. En este caso, se emplea la función parse_number del paquete readr, que es especialmente útil para convertir valores almacenados como texto en datos numéricos. Esta función extrae automáticamente los componentes numéricos de una cadena de texto y los convierte en un tipo de dato numérico, parse_number eliminará los caracteres no numéricos, como letras, símbolos o comas Este proceso es clave para trabajar con columnas que combinan datos numéricos y caracteres en un mismo campo, ya que permite su correcta utilización en operaciones matemáticas y estadísticas.

Adicionalmente, se utiliza la función mutate del paquete dplyr, la cual es útil para modificar o crear nuevas columnas en un conjunto de datos. Dentro de mutate, se aplica parse_number a las columnas seleccionadas para asegurar su transformación. Esto permite trabajar de forma eficiente con múltiples variables en un solo paso, sin necesidad de manipularlas de forma individual.

```
colnames(DATA)
str(DATA)
DATA_Manipulada <- DATA %>%
  mutate(
    # Conversión de columnas a numérico
    Growth = parse_number(Growth,
```

```
                      locale = locale(decimal_mark = ".")),
  Ln_Inflation = parse_number(Ln_Inflation,
                              locale = locale(decimal_mark = ".")),
  Ln_Seniority = parse_number(Ln_Seniority,
                              locale = locale(decimal_mark = ".")),
  Ln_Fixed_Assets = parse_number(Ln_Fixed_Assets,
                                 locale = locale(decimal_mark = ".")),
  Ln_Current_Assets = parse_number(Ln_Current_Assets,
                                   locale = locale(decimal_mark = ".")),
  Ln_Inventory = parse_number(Ln_Inventory,
                              locale = locale(decimal_mark = ".")),
  Ln_Receivables = parse_number(Ln_Receivables,
                                locale = locale(decimal_mark = ".")),
  Ln_Other_Assets = parse_number(Ln_Other_Assets,
                                 locale = locale(decimal_mark = ".")),
  Ln_Cash = parse_number(Ln_Cash, locale =
                           locale(decimal_mark = ".")),
  Ln_Total_Assets = parse_number(Ln_Total_Assets,
                                 locale = locale(decimal_mark = ".")),
  Liquidity1 = parse_number(Liquidity1,
                            locale = locale(decimal_mark = ".")),
  Liquidity1_Dummy = parse_number(Liquidity1_Dummy,
                                  locale = locale(decimal_mark = ".")),
  Ln_Non_Current_Liabilities = parse_number
  (Ln_Non_Current_Liabilities,
    locale = locale(decimal_mark = ".")),
  Ln_Current_Liabilities = parse_number(Ln_Current_Liabilities,
                                        locale = locale(decimal_mark = ".")),
  Ln_Total_Liabilities = parse_number(Ln_Total_Liabilities,
                                      locale = locale(decimal_mark = ".")),
  Ln_Equity = parse_number(Ln_Equity,
                           locale = locale(decimal_mark = ".")),
  ROE = parse_number(ROE, locale =
                       locale(decimal_mark = ".")),
  ROA = parse_number(ROA,
                     locale = locale(decimal_mark = ".")),
  Collection_Period = parse_number(Collection_Period,
                                   locale = locale(decimal_mark = ".")),
  Credit_Period = parse_number(Credit_Period,
                               locale = locale(decimal_mark = ".")),
  ROEE = parse_number(ROEE, locale =
                        locale(decimal_mark = ".")),
  ROAA = parse_number(ROAA,
                      locale = locale(decimal_mark = ".")),
  CollectionPeriod = parse_number(CollectionPeriod,
                                  locale = locale(decimal_mark = ".")),
  Payment_Period = parse_number(Payment_Period,
                                locale = locale(decimal_mark = ".")),
  PMC_PMP = parse_number(PMC_PMP,
                         locale = locale(decimal_mark = ".")),
  Net_Asset_Turnover = parse_number(Net_Asset_Turnover,
                                    locale = locale(decimal_mark = ".")),
  Inventory_Turnover = parse_number(Inventory_Turnover,
```

```
                                    locale = locale(decimal_mark = ".")),
 Solvency_Turnover = parse_number(Solvency_Turnover,
                                   locale = locale(decimal_mark = ".")),
 Asset_Turnover = parse_number(Asset_Turnover,
                               locale = locale(decimal_mark = ".")),
 InventoryTurnover = parse_number(InventoryTurnover,
                                   locale = locale(decimal_mark = ".")),
 SolvencyTurnover = parse_number(SolvencyTurnover,
                                  locale = locale(decimal_mark = ".")),
 Liquidity_Ratio = parse_number(Liquidity_Ratio,
                                locale = locale(decimal_mark = ".")),
 Leverage = parse_number(Leverage, locale =
                           locale(decimal_mark = ".")),
 Profit_per_Employee = parse_number(Profit_per_Employee,
                                    locale = locale(decimal_mark = ".")),
 Operating_Revenue_per_Employee = parse_number
 (Operating_Revenue_per_Employee, locale = locale(decimal_mark = ".")),
 Levera = parse_number(Levera,
                       locale = locale(decimal_mark = ".")),
 Profit_Employee = parse_number(Profit_Employee,
                                locale = locale(decimal_mark = ".")),
 OperatingRevenue_Employee = parse_number
 (OperatingRevenue_Employee, locale = locale(decimal_mark = ".")),
 Shareholder_Direct_Percentage = parse_number
 (Shareholder_Direct_Percentage, locale = locale(decimal_mark = ".")),
 Shareholder_Total_Percentage = parse_number
 (Shareholder_Total_Percentage, locale = locale(decimal_mark = ".")),
 CSH_Direct_Percentage = parse_number(CSH_Direct_Percentage,
                                      locale = locale(decimal_mark = ".")),

 # Conversión de columnas a Date
 Incorporation_Date = as.Date(Incorporation_Date, origin = "1899-12-30"),
 End_Date = as.Date(End_Date, origin = "1899-12-30"),

# Conversión de columnas a character
 Country = as.character(Country),
 Identity = as.character(Identity),
 Legal_Form = as.character(Legal_Form),
 Legal_Form_Tabul = as.character(Legal_Form_Tabul),

#Conversion de columnas a factor
 ADV_Gender= parse_factor(ADV_Gender,
                          levels = c("M","F","M\nM","M\nM\nM","M\nF"),
                          ordered = TRUE),
 BvD_Independence_Indicator= parse_factor(BvD_Independence_Indicator,
                                          levels = c("A+", "A", "A-",
                                                     "B+", "B", "B-", "C+",
                                                     "C", "C-", "D", "U"),
                                          ordered = TRUE),
 Standard_Legal_Form = parse_factor(Standard_Legal_Form,
                                    levels = c("Public limited companies",
                                               "Private limited companies",
                                               "Partnerships",
```

```
                                                    "Other legal forms"),
                                        ordered = FALSE))
```

10 Descripción de las variables

10.1 Tabla 1

Esta investigación se ha centrado en examinar los factores que determinan, en mayor medida, la rentabilidad de las empresas que operan en el sector de piedras naturales. Para el estudio del rendimiento de las empresas, las variables más comúnmente utilizadas son ROA (rentabilidad sobre activos—rentabilidad económica) y ROE (rentabilidad sobre el capital—rentabilidad financiera), debido a su capacidad para medir las inversiones en términos de activos y patrimonio. Investigaciones previas han utilizado ROA y ROE como variables dependientes. Por lo tanto, en los dos análisis empíricos implementados en este estudio, ROA y ROE fueron consideradas como variables dependientes. Ambas son variables cuantitativas continuas. Las variables explicativas bajo análisis pueden agruparse en tres categorías distintas:

- Variables asociadas a la empresa, como la edad de la empresa en el mercado, así como variables financieras como el volumen de activos, apalancamiento, ingreso operativo total, rotación de existencias, rotación de activos, periodo promedio de cobro, periodo promedio de pago, crecimiento de la empresa, y la forma jurídica.
- Variables asociadas al entorno económico, como el país, el producto interno bruto (PIB), y el nivel de inflación.
- Variables vinculadas a la diversidad en la gestión empresarial, identificada en esta investigación por la variable de género.

Las siguientes variables independientes fueron consideradas en el estudio: tamaño de la empresa, medido por el volumen de activos; el grado de apalancamiento; el crecimiento de la empresa; el cambio en el PIB del país; la inflación; y el porcentaje de directores de la junta femenina. Finalmente, las siguientes variables fueron utilizadas como variables de control: ingreso operativo, rotación de existencias, rotación de activos, periodos promedio de cobro y pago, edad de la empresa, y forma jurídica.

`Rentabilidad financiera (ROE)`

El indicador ROE expresa la capacidad de una empresa para generar ganancias a través de un uso productivo de las contribuciones de los accionistas y una gestión eficiente. Se calcula como la relación entre el beneficio neto después de impuestos y el patrimonio de los accionistas. 3.2.2. Rentabilidad económica (ROA) ROA se define como el beneficio neto después de impuestos dividido por el total de los activos, y también ha sido ampliamente utilizado en la literatura.

`Tamaño de la empresa`

El tamaño de la empresa es considerado un factor importante al explicar la rentabilidad. La teoría sugiere que las empresas más grandes tienen mayor acceso a los mercados financieros y obtienen mejores tasas de interés al explotar las economías de escala. Según algunos estudios, el tamaño de la empresa tiene un efecto positivo y significativo sobre la rentabilidad.

`Endeudamiento`

El endeudamiento es uno de los factores más críticos al analizar el rendimiento corporativo. El ratio de deuda se define como el total de deudas de una empresa dividido por sus activos totales. Los resultados de investigaciones anteriores generalmente encuentran una relación inversa entre el nivel de endeudamiento de una empresa y su rentabilidad.

`Crecimiento`

Algunas investigaciones previas utilizaron el crecimiento de la empresa como el cambio porcentual en los ingresos operativos. Sin embargo, como en este estudio, otros investigadores consideran el crecimiento como el cambio porcentual en los activos totales, mostrando una relación positiva y significativa entre el cambio porcentual en los activos y la rentabilidad de la empresa.

`Variación del PIB`

El PIB es uno de los indicadores más utilizados para medir la actividad económica dentro de un país. El crecimiento económico refleja las condiciones macroeconómicas generales. Se presume que un cambio en el PIB puede influir en el rendimiento de las empresas, con una relación positiva esperada entre la variación del PIB y la rentabilidad.

`Inflación`

El índice de inflación es otro indicador macroeconómico comúnmente utilizado en estudios de rentabilidad. El efecto de la inflación en la rentabilidad de una empresa depende de si la inflación es anticipada o no. Se espera que haya una relación negativa entre la inflación no anticipada y la rentabilidad de las empresas.

`Género`

Muchos estudios han analizado la influencia de la diversidad de género en las juntas directivas sobre la rentabilidad de las empresas, encontrando que una mayor presencia femenina tiene una influencia positiva en la rentabilidad.

`Ingreso operativo`

El ingreso operativo de una empresa es considerado un indicador clave de muchos aspectos positivos que apoyan tanto el crecimiento como la rentabilidad. Este estudio utiliza el logaritmo natural del ingreso operativo para determinar su relación con la rentabilidad de la empresa.

`Rotación de existencias`

El ratio de rotación de existencias es una medida importante para evaluar la eficiencia de la gestión de inventarios. Una alta rotación de existencias generalmente indica una gestión eficiente. Se espera que esta relación sea positiva con la rentabilidad.

`Rotación de activos`

La rotación de activos mide la eficiencia de una empresa en el uso de sus activos para generar ingresos operativos. Los resultados previos muestran resultados mixtos, con algunos estudios que indican que la rotación de activos tiene un impacto positivo en la rentabilidad.

`Periodo promedio de cobro`

El periodo promedio de cobro representa el número promedio de días que una empresa tarda en cobrar después de una venta a crédito. En algunos estudios, se encuentra que existe una relación positiva significativa entre el periodo promedio de cobro y la rentabilidad, mientras que otros autores encuentran una relación negativa.

`Periodo promedio de pago`

Este indicador muestra el número promedio de días que una empresa tarda en pagar sus deudas a corto plazo. Estudios empíricos han encontrado una relación negativa entre el periodo promedio de pago y la rentabilidad de la empresa.

`Edad`

La edad de la empresa, definida como el número de años que ha estado en el mercado, tiene una relación compleja con la rentabilidad. Mientras que algunos estudios encuentran una relación positiva significativa, otros argumentan que la edad tiene un efecto negativo sobre la rentabilidad.

`Forma jurídica`

La forma jurídica ha sido utilizada para analizar la rentabilidad, particularmente en empresas de responsabilidad limitada. Sin embargo, algunos estudios no han encontrado una relación significativa entre la forma jurídica y la rentabilidad.

`País`

El país en el que se ubica la empresa puede influir significativamente en su rentabilidad, ya que las condiciones económicas y regulatorias varían según la región.

```
library(tibble)

# Crear la tabla de variables con saltos de línea
tabla_esp <- tibble::tibble(
  Abreviatura = c("ROE", "ROA", "Size", "Debt", "Growth", "VarGDP", "Inflat",
                  "Gender", "OpInc", "StockT", "AssetT", "ARP", "APP",
                  "Age", "LForm", "Country"),
  Variable = c("Retorno sobre el patrimonio", "Retorno sobre los activos",
               "Tamaño de la empresa", "Endeudamiento",
               "Crecimiento de la empresa", "Cambio en el PIB",
               "Inflación del país", "Diversidad de género",
               "Ingreso operativo", "Rotación del inventario",
               "Rotación de activos", "Período promedio de cobro",
               "Período promedio de pago", "Edad de la empresa",
               "Forma jurídica", "País de residencia de la empresa"),
  Definición = c(
    "Beneficio neto dividido por el patrimonio",
    "Beneficio neto dividido por los activos totales",
    "Logaritmo natural del total de activos de la empresa",
    "Pasivos totales divididos por los activos totales",
    "Porcentaje de cambio en los activos totales",
    "Porcentaje de cambio en el PIB",
    "Inflación del país en el año",
    "Porcentaje de mujeres en el consejo de administración",
    "Logaritmo natural del ingreso operativo",
```

```
    "Costo de ventas dividido por inventario",
    "Ingreso operativo dividido por activos totales",
    "Promedio de días que la empresa tarda en recibir pagos",
    "Promedio de días que la empresa tarda en pagar a proveedores",
    "Edad de la empresa en años",
    "La forma jurídica de la empresa: Sociedad anónima,
    Sociedad limitada, Cooperativa, Otras formas legales",
    "Variable igual a 1 para España y 0 para Italia"
  )
)
```

```
# Crear la tabla en gt()
tabla_esp %>%
  gt() %>%
  tab_header(
    title = md("**Tabla 1. Definición de variables (Español)**")
  ) %>%
  cols_label(
    Abreviatura = md("**Abreviatura**"),
    Variable = md("**Variable**"),
    Definición = md("**Definición**")
  ) %>%
  cols_width(
    Definición ~ px(300)  # Ajusta el ancho de la columna Definición
  ) %>%
  tab_options(
    column_labels.font.size = px(12),
    table.font.size = px(12),
    data_row.padding = px(5),
    row.striping.include_table_body = TRUE
  ) %>%
  tab_style(
    style = list(
      cell_text(align = "center")
    ),
    locations = cells_title(groups = "title")
  )
```

Creacion de varables faltantes

En esta sección, se calculan nuevas variables derivadas a partir de los datos existentes en el conjunto DATA_Manipulada. Estas transformaciones tienen como objetivo enriquecer el análisis mediante indicadores adicionales que permiten evaluar distintos aspectos financieros y operativos de las entidades analizadas. Entre los cálculos realizados, se incluyen:

```
·Debt (Endeudamiento):Total de pasivos dividido por el total de activos
    - Debt = Total_Liabilities / Total_Assets
·Oplnc    : Logaritmo natural de los ingresos de explotación
    - OpInc= log(Operating_Revenue)
·StockT(Rotación de inventarios): Ingresos de explotación dividido
por el inventario
    - AssetT = Operating_Revenue / Total_Assets
·AssetT   (Rotación de activos): Ingresos de explotación dividido
por el inventario
```

Tabla 1. Definición de variables (Español)

Abreviatura	Variable	Definición
ROE	Retorno sobre el patrimonio	Beneficio neto dividido por el patrimonio
ROA	Retorno sobre los activos	Beneficio neto dividido por los activos totales
Size	Tamaño de la empresa	Logaritmo natural del total de activos de la empresa
Debt	Endeudamiento	Pasivos totales divididos por los activos totales
Growth	Crecimiento de la empresa	Porcentaje de cambio en los activos totales
VarGDP	Cambio en el PIB	Porcentaje de cambio en el PIB
Inflat	Inflación del país	Inflación del país en el año
Gender	Diversidad de género	Porcentaje de mujeres en el consejo de administración
OpInc	Ingreso operativo	Logaritmo natural del ingreso operativo
StockT	Rotación del inventario	Costo de ventas dividido por inventario
AssetT	Rotación de activos	Ingreso operativo dividido por activos totales
ARP	Período promedio de cobro	Promedio de días que la empresa tarda en recibir pagos
APP	Período promedio de pago	Promedio de días que la empresa tarda en pagar a proveedores
Age	Edad de la empresa	Edad de la empresa en años
LForm	Forma jurídica	La forma jurídica de la empresa: Sociedad anónima, Sociedad limitada, Cooperativa, Otras formas legales
Country	País de residencia de la empresa	Variable igual a 1 para España y 0 para Italia

```
      - StockT = Operating_Revenue / Inventory
  ·Age(Antigüedad) :Edad de la empresa en años
      - Age = Seniority / 365
```

```
DATA_Manipulada<- DATA_Manipulada %>%
  mutate(Debt = Total_Liabilities / Total_Assets,
         OpInc= log(Operating_Revenue),
         AssetT = Operating_Revenue / Total_Assets ,
         StockT = Operating_Revenue / Inventory,
         Age = Seniority / 365 )


#selección de variables(existentes en la DATA) a utilizar para la Tabla 1
DATAM_SELECT <- DATA_Manipulada %>%
  select(ROE, ROA, Ln_Total_Assets,Debt, Growth, GDP_Var, Inflation, Gender,
         OpInc,StockT, AssetT, CollectionPeriod,
         Payment_Period, Age, Legal_Form, Country) %>%
  rename(
    Size = Ln_Total_Assets,
    VarGDP = GDP_Var,
    Inflat = Inflation,
    ARP = CollectionPeriod,
    APP = Payment_Period,
    LForm = Legal_Form)


#aplica a todas las columnas numéricas, redondeando los valores a dos decimales.
DATA_SELECT <- DATAM_SELECT %>%
  mutate(across(where(is.numeric), ~ round(.x, 2)))
```

SubDataset

Para facilitar el análisis comparativo, se procedió a dividir el conjunto de datos principal (DATA_SELECT) en tres subconjuntos, basados en la variable Country, que identifica la procedencia de las observaciones:

- GENERAL: Incluye todas las observaciones del conjunto original, excluyendo las variables Country y LForm, ya que no son necesarias para este análisis en particular. Este subconjunto representa una visión global del dataset.
- ESPAÑA: Contiene únicamente las observaciones donde el valor de Country es "1", correspondiente a España. Además, se eliminan las columnas Country y LForm, manteniendo únicamente las variables relevantes para el análisis específico de este país.
- ITALIA: Comprende las observaciones donde el valor de Country es "0", correspondiente a Italia, siguiendo el mismo criterio de selección y limpieza aplicado en el subconjunto de España.

```
#FILTAR OBSERVACION SEGUN EL PAIS
OB_ES<-DATA_SELECT %>% filter(Country=="1") #ESPAÑA
OB_IT<-DATA_SELECT %>% filter(Country=="0") #ITALIA

GENERAL <- DATA_SELECT %>% select (-c (Country, LForm))
ESPAÑA <-OB_ES %>% select (-c (Country, LForm))
ITALIA<- OB_IT %>% select (-c (Country, LForm))
```

11 Estadísticas descriptivas y correlaciones

11.1 Tabla 2

La Tabla 2 presenta las principales estadísticas descriptivas de las variables utilizadas para explicar la rentabilidad de las empresas que operan en el sector de la piedra natural en España e Italia durante el período 2015-2019. Los datos se presentan tanto para la muestra total como para cada submuestra según el país donde se ubica la empresa. Además, se incluye la prueba de medias, la cual muestra diferencias significativas entre ambos países.

Para empezar se reemplazan los valores infinitos por NA para evitar sesgos en los cálculos. Posteriormente, se generan resúmenes estadísticos que incluyen la media, desviación estándar, valores mínimos y máximos de cada variable. Esto permite obtener una visión clara de las principales características de los datos en cada conjunto. Finalmente, las tablas son unidas en diferentes combinaciones para su análisis comparativo.

```
G <- GENERAL %>%
  mutate(across(everything(), ~ ifelse(is.infinite(.), NA, .)))
RE_G<- G %>% get_summary_stats() %>% select(variable, mean, sd, min, max)


E <- ESPAÑA %>%
  mutate(across(everything(), ~ ifelse(is.infinite(.), NA, .)))
RE_E <- E %>% get_summary_stats() %>% select(variable, mean, sd, min, max)

I <- ITALIA %>%
  mutate(across(everything(), ~ ifelse(is.infinite(.), NA, .)))
RE_I<- I %>% get_summary_stats() %>% select(variable, mean, sd, min, max)

RE_UNIDO_E_I <- full_join(RE_E, RE_I, by = "variable")
```

```r
RE_UNIDO <- RE_G %>%
  full_join(RE_E, by = "variable") %>%
  full_join(RE_I, by = "variable")

RE_UNIDO <- RE_G %>%
  left_join(RE_E, by = "variable") %>%
  left_join(RE_I, by = "variable")
```

se calcula el valor p (p-value) para cada variable mediante una prueba t de Student, comparando las muestras de España e Italia. Esto permite evaluar si existen diferencias estadísticamente significativas entre las dos poblaciones para cada variable. Además, se agrega una columna de significancia con asteriscos que identifican el nivel de significancia ($p < 0.10$, $p < 0.05$, $p < 0.01$). Esto facilita la interpretación de los resultados en términos de relevancia estadística.

```r
# Calcular p-value para cada variable
p_values <- map_dfr(RE_E$variable, ~ {
  var <- .x
  t_test <- t.test(E[[var]], I[[var]], var.equal = TRUE)
  tibble(variable = var, p_value = round(t_test$p.value, 2))
})
# Agregar una columna de significancia con asteriscos
p_values <- map_dfr(RE_E$variable, ~ {
  var <- .x
  t_test <- t.test(E[[var]], I[[var]], var.equal = TRUE)
  p_value <- round(t_test$p.value, 4)

  tibble(
    variable = var,
    p_value = paste0(
      format(p_value, nsmall = 4, scientific = FALSE),
      case_when(
        p_value < 0.01 ~ "***",
        p_value < 0.05 ~ "**",
        p_value < 0.10 ~ "*",
        TRUE ~ ""
      )
    )
  )
})
```

se combinan los resúmenes estadísticos de España e Italia en una sola tabla, añadiendo los valores p obtenidos previamente. Luego, esta tabla es fusionada con las estadísticas descriptivas del conjunto general. El objetivo es consolidar toda la información en una única tabla que permita comparar fácilmente las estadísticas generales, españolas e italianas, junto con los resultados de las pruebas de diferencias de medias.

```r
# Combinar las estadísticas descriptivas de España e Italia
spain_italy_table <- RE_E %>%
  left_join(RE_I, by = "variable", suffix = c("_spain", "_italy"))

# Agregar los p-values a la tabla combinada de España e Italia
spain_italy_table <- spain_italy_table %>%
  left_join(p_values, by = "variable")
```

```r
# Combinar la tabla general (RE_G) con la tabla de España e Italia
tabla_2 <- RE_G %>%
  left_join(spain_italy_table, by = "variable", suffix = c("_total", ""))
```

Finalmente, en esta sección, se formatea la tabla combinada utilizando la librería gt para generar una presentación clara y profesional. Se ajustan los números a cuatro decimales, se renombran las columnas para mayor claridad y se agrupan las estadísticas bajo encabezados representativos ("Total Sample", "Spain", "Italy"). Asimismo, se agrega un encabezado para la tabla, resaltando que se trata de estadísticas descriptivas y pruebas de diferencias de medias por país. Este formato facilita la interpretación visual de los resultados.

```r
# Formatear la tabla con gt()
tabla_2 %>%
  gt() %>%
  fmt_number(
    columns = c(mean, sd, min, max,
                mean_spain, sd_spain, min_spain, max_spain,
                mean_italy, sd_italy, min_italy, max_italy),
    decimals = 2
  ) %>%
cols_label(
  variable = md("**Variable**"),
  mean = md("**M**"),
  sd = md("**SD**"),
  min = md("**Min**"),
  max = md("**Max**"),
  mean_spain = md("**M**"),
  sd_spain = md("**SD**"),
  min_spain = md("**Min**"),
  max_spain = md("**Max**"),
  mean_italy = md("**M**"),
  sd_italy = md("**SD**"),
  min_italy = md("**Min**"),
  max_italy = md("**Max**"),
  p_value = md("**valor p**")
) %>%
  tab_spanner(
    label = md("**Muestra Total**"),
    columns = c(mean, sd, min, max)
  ) %>%
  tab_spanner(
    label = md("**España**"),
    columns = c(mean_spain, sd_spain, min_spain, max_spain)
  ) %>%
  tab_spanner(
    label = md("**Italia**"),
    columns = c(mean_italy, sd_italy, min_italy, max_italy)
  ) %>%
  tab_header(
    title = md("**Tabla 2. Estadísticas descriptivas y prueba de
                diferencia de medias por país**")
  )%>%
  tab_options(
    column_labels.font.size = px(7), # Tamaño del texto de encabezado
```

Tabla 2. Estadísticas descriptivas y prueba de diferencia de medias por país

	Muestra Total				España				Italia				
Variable	M	SD	Min	Max	M	SD	Min	Max	M	SD	Min	Max	valor p
ROE	4.86	50.63	-973.58	169.28	2.81	55.56	-920.19	164.48	6.47	46.38	-973.58	169.28	0.0992*
ROA	2.99	8.33	-79.25	65.44	2.96	8.79	-79.25	65.44	3.02	7.95	-65.89	46.23	0.8717
Size	8.25	1.10	1.10	12.63	8.25	1.26	1.10	12.63	8.25	0.95	2.75	12.26	0.9788
Debt	0.54	0.31	0.00	2.59	0.44	0.30	0.00	1.86	0.61	0.30	0.00	2.59	0.0000***
Growth	0.64	15.68	-1.00	479.78	0.66	17.00	-1.00	479.78	0.62	14.55	-1.00	454.58	0.9599
VarGDP	1.82	1.07	0.02	3.84	2.83	0.65	0.02	3.84	0.99	0.47	0.29	1.67	0.0000***
Inflat	0.65	0.77	-0.50	1.96	0.73	0.98	-0.50	1.96	0.59	0.54	-0.09	1.23	0.0000***
Gender	21.38	31.00	0.00	100.00	24.36	30.47	0.00	100.00	18.95	31.22	0.00	100.00	0.0000***
OpInc	7.48	1.30	-1.82	12.43	7.55	1.26	1.27	12.43	7.42	1.33	-1.82	11.92	0.0171**
StockT	66.77	353.73	-0.15	9,680.32	47.80	254.31	0.03	3,647.51	82.09	416.62	-0.15	9,680.32	0.0387**
AssetT	0.69	0.70	-0.06	15.20	0.77	0.91	0.00	15.20	0.63	0.46	-0.06	3.69	0.0000***
ARP	132.06	140.49	0.00	981.75	132.71	130.10	0.00	967.13	131.52	148.53	0.00	981.75	0.8432
APP	64.92	104.16	0.00	993.43	47.50	77.82	0.00	993.43	79.21	119.74	0.00	962.24	0.0000***
Age	27.84	17.11	-2.95	103.41	26.45	12.73	-2.95	62.05	28.96	19.91	-2.55	103.41	0.0005***

```
    table.font.size = px(7) # Tamaño del texto en la tabla
  )%>%
  tab_style(
    style = list(
      cell_text(align = "center")
    ),
    locations = cells_title(groups = "title")
  )
```

11.2 Tabla 3

La Tabla 3 muestra los resultados de la prueba de chi-cuadrado aplicada a la forma legal según el país de referencia. Se encuentran diferencias significativas entre España e Italia. Mientras que en España hay muchas empresas con la forma legal de sociedad anónima, en Italia se observa una clara preferencia por las sociedades de responsabilidad limitada. Del mismo modo, Italia cuenta con cooperativas y otras formas sociales en este sector, las cuales prácticamente no existen en España. De hecho, España es un país donde la forma legal de cooperativa es ampliamente utilizada en otros sectores, como la agricultura, pero está casi ausente en el sector de la piedra natural.

Se tiene como propósito contar las frecuencias de los niveles de la variable categórica LForm (que representa la forma legal de las empresas) en diferentes conjuntos de datos.

En primer lugar, se trabaja con el conjunto de datos DATA_SELECT, que contiene información general. A través de la función count(LForm), se calcula el número total de empresas para cada nivel de la variable LForm. Posteriormente, se renombra la columna que contiene los conteos como "Total sample" para indicar que representa la totalidad de la muestra.

```
LF_GENERAL <- DATA_SELECT %>% count(LForm) %>% rename("Total sample" = n)
LF_ESPAÑA<- OB_ES %>% count(LForm) %>% rename(Spain = n)
LF_ITALIA<- OB_IT %>% count(LForm) %>% rename(Italy = n)
```

Se utiliza la función full_join para unir tres tablas basadas en una columna común. La unión completa asegura que todas las observaciones de las tablas se conserven, incluso si no hay coincidencias entre ellas. Luego, se usa is.na() para identificar y reemplazar los valores faltantes con ceros en el resultado de la unión.

```
# Unir las tablas por 'LForm' (el tipo de empresa)
T_LEGAL_FORM <- LF_GENERAL %>%
  full_join(LF_ESPAÑA, by = "LForm") %>%
  full_join(LF_ITALIA, by = "LForm")
```

Posteriormente, se emplean las funciones distinct para obtener los valores únicos de dos columnas relacionadas con las formas legales. Se define un vector que asigna nombres descriptivos a códigos numéricos representando tipos de empresas. Finalmente, la función mutate se usa para reemplazar los códigos numéricos por los nombres de los tipos de empresas, y rename cambia el nombre de la columna LForm a "Legal form" para una mejor comprensión.

```
# Rellenar con ceros en caso de valores faltantes
T_LEGAL_FORM[is.na(T_LEGAL_FORM)] <- 0

DATA_Manipulada %>% distinct(Standard_Legal_Form)
DATA_Manipulada %>% distinct(Legal_Form)
tipo_empresa <- c("0" = "Public limited company",
                  "1" = "Private limited company",
                  "3" = "Cooperative",
                  "4" = "Other legal forms")
# Reemplazar los códigos numéricos con los nombres de tipos de empresas
T_LEGAL_FORM <-T_LEGAL_FORM %>%
  mutate(LForm = tipo_empresa[as.character(LForm)])%>%
  rename("Legal form" = LForm)
```

Se realiza el calculo del test de Chi-cuadrado para comparar las frecuencias observadas entre dos columnas específicas (en este caso, España e Italia). Primero, selecciona los valores de estas dos columnas y los convierte en una matriz mediante as.matrix(). Luego, se aplica la función chisq.test() para calcular el test de Chi-cuadrado sobre las frecuencias observadas.

```
#Calculo de Chi-squared test
observed <- T_LEGAL_FORM %>%
  select(Spain, Italy) %>%
  as.matrix()

chisq_test <- chisq.test(observed)

# Extraer los valores del test y formatearlos
chi_squared_value <- round(chisq_test$statistic, 4)

# Formatear p-value manualmente (sin "<")
p_value <- ifelse(chisq_test$p.value < 0.0000, "0.0000",
                  sprintf("%.4f", chisq_test$p.value))

# Agregar el resultado del Chi-cuadrado a la tabla
tabla_3 <- T_LEGAL_FORM %>%
  mutate(`Chi-squared test` = ifelse(row_number() == 1,
                                     paste0(chi_squared_value,
                                            " (", p_value, ")"),
                                     ""))
```

Finalmente se utiliza la función gt() para crear una tabla estilizada, personalizándola con opciones para ajustar títulos, etiquetas de columnas, alineación de celdas, bordes y notas al pie.

La Tabla 3 muestre los resultados del test de Chi-cuadrado junto con las frecuencias de las formas legales por país, mejorando así la presentación de los datos y facilitando su comprensión.

Tabla 3. Forma jurídica por país

Forma jurídica	Muestra total	España	Italia	Prueba de chi-cuadrado
Public limited company	405	325	80	272.8377 (0.0000)
Private limited company	1765	670	1095	
Cooperative	65	20	45	
Other legal forms	35	0	35	

```
# Formatear la tabla con gt()
tabla_3 %>%
  gt() %>%
  tab_header(
    title = md("**Tabla 3. Forma jurídica por país**"),
    subtitle = ""
  ) %>%
  cols_label(
    `Legal form` = md("**Forma jurídica**"),
    `Total sample` = md("**Muestra total**"),
    Spain = md("**España**"),
    Italy = md("**Italia**"),
    `Chi-squared test` = md("**Prueba de chi-cuadrado**")
  ) %>%
  cols_align(
    align = "left",
    columns = `Legal form`
  ) %>%
  cols_align(
    align = "center",
    columns = c(`Total sample`, Spain, Italy, `Chi-squared test`)
  )%>%
  tab_options(
    column_labels.font.size = px(12), # Tamaño del texto de encabezado
    table.font.size = px(12) # Tamaño del texto en la tabla
  )%>%
  tab_style(
    style = list(
      cell_text(align = "center")
    ),
    locations = cells_title(groups = "title")
  )
```

11.3 Tabla 4

Muestra la matriz de correlación de Pearson entre las variables continuas utilizadas en el estudio empírico. Se puede observar que no existen correlaciones elevadas entre los regresores que puedan generar problemas de colinealidad en el análisis multivariante posterior.

Además, todos los regresores, excepto la inflación, presentan una correlación significativa con las variables dependientes (ROE y ROA). Específicamente, el volumen de activos, los períodos promedio de cobro y pago, la antigüedad de la empresa, el cambio en el PIB y el género de la dirección están negativamente correlacionados con ROE y ROA. Para el cambio en el PIB y el género, la relación solo es significativa en el caso del ROE.

Tabla 4. Correlación pearson para las variables continuas

	ROE	ROA	Size	Debt	Growth	VarGDP	Inflat	Gender	OpInc	StockT	AssetT	ARP	APP	Age
ROE														
ROA	0.44****													
Size	-0.064**	-0.064**												
Debt	-0.07**	-0.24****	-0.2****											
Growth	0.02	0.059*	-0.012	-0.013										
VarGDP	-0.025	-0.0083	-0.02	-0.22****	0.0043									
Inflat	-0.024	0.0041	0.028	-0.033	-0.00053	-0.055**								
Gender	0.0084	-0.012	-0.048*	-0.046*	-0.024	0.072***	0.0076							
OpInc	0.071**	0.2****	0.58****	-0.1****	0.0028	0.013	0.038	-0.078***						
StockT	0.068**	0.12****	-0.011	0.018	0.018	-0.053*	-0.032	-0.0082	0.067**					
AssetT	0.13****	0.3****	-0.43****	0.15****	0.019	0.083***	0.0062	-0.029	0.21****	0.11****				
ARP	-0.07**	-0.17****	0.09****	-0.022	0.012	0.0094	-0.011	-0.031	-0.16****	-0.0067	-0.25****			
APP	-0.1****	-0.19****	-0.034	0.2****	-0.0042	-0.13****	-0.031	-0.051*	-0.2****	-0.033	-0.15****	0.32****		
Age	-0.0053	-0.061**	0.32****	-0.17****	-0.06*	-0.036	-0.056**	0.07***	0.14****	-0.039	-0.21****	0.034	-0.045*	

Por otro lado, la correlación es positiva para las variables que miden el ingreso operativo, la rotación de inventario, la rotación de activos y el crecimiento, aunque solo la correlación entre ROA y crecimiento es significativa.

En cuanto al endeudamiento, la correlación es significativa y negativa para el ROA, pero positiva para el ROE.

Para calcular la matriz de correlación, primero se usa select(), que permite extraer únicamente las variables numéricas de interés. Luego, con cor_mat(method = "pearson"), se calcula la matriz de correlación utilizando el método de Pearson, que mide la relación lineal entre las variables.

Posteriormente, cor_mark_significant() se encarga de marcar aquellas correlaciones que son estadísticamente significativas, lo que ayuda a interpretar los resultados.

```
# Generar la matriz de correlación
correlation_matrix <- G %>%
  select(ROE, ROA, Size, Debt, Growth, VarGDP, Inflat, Gender, OpInc, StockT,
         AssetT, ARP, APP, Age) %>%
  cor_mat(method = "pearson") %>%
  cor_mark_significant() # Añade los niveles de significancia
```

Finalmente, para visualizar los datos de manera clara, se utiliza gt(), que convierte la matriz en una tabla interactiva, y tab_options(), que ajusta el tamaño del texto de la tabla para mejorar la presentación.

```
# Mostrar la matriz de correlación en una vista interactiva
correlation_matrix  %>%
  gt() %>%
  tab_header(
    title = md("**Tabla 4. Correlación pearson para las variables continuas**")
  ) %>%
  tab_options(
    column_labels.font.size = px(5), # Tamaño del texto de encabezado
    table.font.size = px(5) # Tamaño del texto en la tabla
  )%>%
  tab_style(
    style = list(
      cell_text(align = "center")
    ),
    locations = cells_title(groups = "title")
  )
```

11.4 Tabla 5

Muestra los resultados del análisis de varianza del ROA y ROE en función de la forma jurídica de la empresa. Se observa que, efectivamente, existe una relación significativa entre la forma jurídica de la empresa y los

dos tipos de rentabilidad, tanto en España como en Italia.

Se calcula el promedio de los indicadores financieros ROE (Return on Equity) y ROA (Return on Assets) según la forma legal de las empresas. Este análisis se realiza de manera general, considerando toda la muestra de datos, y luego se repite específicamente para empresas en España e Italia. Para ello, se utiliza la función group_by() para agrupar los datos según la forma legal (LForm), seguida de summarise() para calcular la media de ROE y ROA, excluyendo valores nulos mediante na.rm = TRUE. Estos cálculos permiten obtener una visión comparativa del desempeño financiero según la estructura legal de las empresas en diferentes regiones.

```
# Calcular promedios de ROE y ROA por tipo de empresa (Legal Form) GENRAL
legal_form_means_GENERAL <- DATA_SELECT %>% select(LForm , Country, ROA, ROE)%>%
  group_by(LForm) %>%
  summarise(
    Mean_ROE = mean(ROE, na.rm = TRUE),
    Mean_ROA = mean(ROA, na.rm = TRUE)
  )
# Calcular promedios de ROE y ROA por tipo de empresa (Legal Form) ESPAÑA
legal_form_means_ESPAÑA <- OB_ES %>% select(LForm , Country, ROA, ROE) %>%
  group_by(LForm) %>%
  summarise(
    Mean_ROE = mean(ROE, na.rm = TRUE),
    Mean_ROA = mean(ROA, na.rm = TRUE)
  )
# Calcular promedios de ROE y ROA por tipo de empresa (Legal Form) ITALIA
legal_form_means_ITALIA <- OB_IT %>% select(LForm , Country, ROA, ROE) %>%
  group_by(LForm) %>%
  summarise(
    Mean_ROE = mean(ROE, na.rm = TRUE),
    Mean_ROA = mean(ROA, na.rm = TRUE)
  )
```

Se realiza un análisis de varianza (ANOVA) para evaluar si existen diferencias estadísticamente significativas en los valores de ROA y ROE según la forma legal de las empresas. Este análisis se lleva a cabo en tres niveles: para toda la muestra, para España y para Italia. Se utiliza la función aov(), que aplica un ANOVA unidireccional sobre la variable dependiente (ROA o ROE) en función de la variable categórica LForm. Los resultados de este análisis permitirán determinar si la estructura legal de una empresa influye significativamente en su desempeño financiero en cada contexto geográfico.

```
# ANOVA por forma legal para ROA y ROE (Muestra total)
anova_roa_total <- aov(ROA ~ LForm, data = DATA_SELECT)
anova_roe_total <- aov(ROE ~ LForm, data = DATA_SELECT)
# ANOVA por forma legal para España
anova_roa_es <- aov(ROA ~ LForm, data = OB_ES)
anova_roe_es <- aov(ROE ~ LForm, data = OB_ES)
# ANOVA por forma legal para Italia
anova_roa_it <- aov(ROA ~ LForm, data = OB_IT)
anova_roe_it <- aov(ROE ~ LForm, data = OB_IT)
```

Los resultados de los cálculos anteriores se combinan en una única tabla. Se inicia con la tabla de promedios generales y, utilizando la función left_join(), se integran los resultados específicos de España e Italia. Además, se renombraron las columnas para mejorar la claridad, indicando a qué país o muestra total pertenecen los valores de ROE y ROA. Finalmente, se redondean los valores a dos decimales mediante mutate(across(where(is.numeric), ~ round(.x, 2))), lo que facilita la presentación y comparación de los datos.

```r
# Combinar los datos de means
tabla_means <- legal_form_means_GENERAL %>%
  rename("ROE_Total_sample" = Mean_ROE, "ROA_Total_sample" = Mean_ROA) %>%
  left_join(
    legal_form_means_ESPAÑA %>%
      rename("ROE_Spain" = Mean_ROE, "ROA_Spain" = Mean_ROA),
    by = "LForm"
  ) %>%
  left_join(
    legal_form_means_ITALIA %>%
      rename("ROE_Italy" = Mean_ROE, "ROA_Italy" = Mean_ROA),
    by = "LForm"
  ) %>%
  mutate(across(where(is.numeric), ~ round(.x, 2)))  # Redondear a 2 decimales
```

Se genera un nuevo dataframe que almacena los resultados de los análisis ANOVA. Se construye una tabla donde cada fila representa los valores de F y p-value para ROE y ROA en la muestra total, España e Italia. Para mejorar la interpretación de los resultados, los valores F se presentan junto con una notación de significancia estadística (, ,), indicando niveles del 1%, 5% y 10% respectivamente. Esta notación se agrega utilizando estructuras condicionales ifelse(), que verifican el valor p correspondiente a cada análisis de varianza.

```r
# Crear un dataframe separado para los valores ANOVA
anova_results <- data.frame(
  `Legal Form` = c("Total sample", "Spain", "Italy"),

  # F-value con significancia para ROE
  `F-value (ROE)` = paste0(
    round(c(
      summary(anova_roe_total)[[1]][["F value"]][1],
      summary(anova_roe_es)[[1]][["F value"]][1],
      summary(anova_roe_it)[[1]][["F value"]][1]
    ), 2),
    c(
      ifelse(summary(anova_roe_total)[[1]][["Pr(>F)"]][1] < 0.01, "***",
             ifelse(summary(anova_roe_total)[[1]][["Pr(>F)"]][1] < 0.05, "**",
                    ifelse(summary(anova_roe_total)[[1]][["Pr(>F)"]][1] < 0.10,
                           "*", "")
             )
      ),
      ifelse(summary(anova_roe_es)[[1]][["Pr(>F)"]][1] < 0.01, "***",
             ifelse(summary(anova_roe_es)[[1]][["Pr(>F)"]][1] < 0.05, "**",
                    ifelse(summary(anova_roe_es)[[1]][["Pr(>F)"]][1] < 0.10,
                           "*", "")
             )
      ),
      ifelse(summary(anova_roe_it)[[1]][["Pr(>F)"]][1] < 0.01, "***",
             ifelse(summary(anova_roe_it)[[1]][["Pr(>F)"]][1] < 0.05, "**",
                    ifelse(summary(anova_roe_it)[[1]][["Pr(>F)"]][1] < 0.10,
                           "*", "")
             )
      )
    )
```

```
  ),

  # p-value para ROE
  `p-value (ROE)` = c(
    summary(anova_roe_total)[[1]][["Pr(>F)"]][1],
    summary(anova_roe_es)[[1]][["Pr(>F)"]][1],
    summary(anova_roe_it)[[1]][["Pr(>F)"]][1]
  ),

  # F-value con significancia para ROA
  `F-value (ROA)` = paste0(
    round(c(
      summary(anova_roa_total)[[1]][["F value"]][1],
      summary(anova_roa_es)[[1]][["F value"]][1],
      summary(anova_roa_it)[[1]][["F value"]][1]
    ), 2),
    c(
      ifelse(summary(anova_roa_total)[[1]][["Pr(>F)"]][1] < 0.01, "***",
             ifelse(summary(anova_roa_total)[[1]][["Pr(>F)"]][1] < 0.05, "**",
                    ifelse(summary(anova_roa_total)[[1]][["Pr(>F)"]][1] < 0.10,
                           "*", "")
             )
      ),
      ifelse(summary(anova_roa_es)[[1]][["Pr(>F)"]][1] < 0.01, "***",
             ifelse(summary(anova_roa_es)[[1]][["Pr(>F)"]][1] < 0.05, "**",
                    ifelse(summary(anova_roa_es)[[1]][["Pr(>F)"]][1] < 0.10,
                           "*", "")
             )
      ),
      ifelse(summary(anova_roa_it)[[1]][["Pr(>F)"]][1] < 0.01, "***",
             ifelse(summary(anova_roa_it)[[1]][["Pr(>F)"]][1] < 0.05, "**",
                    ifelse(summary(anova_roa_it)[[1]][["Pr(>F)"]][1] < 0.10,
                           "*", "")
             )
      )
    )
  ),

  # p-value para ROA
  `p-value (ROA)` = c(
    summary(anova_roa_total)[[1]][["Pr(>F)"]][1],
    summary(anova_roa_es)[[1]][["Pr(>F)"]][1],
    summary(anova_roa_it)[[1]][["Pr(>F)"]][1]
  )
)
```

Se genera una fila adicional para la tabla de resultados con los valores F y p-value, facilitando su integración con la tabla de promedios previamente construida. Aquí, se formatean los valores para que cada celda combine el estadístico F con su correspondiente p-value entre paréntesis. Se hace uso de formatC() para mantener un formato de cuatro decimales en los p-values, asegurando precisión en la presentación de los resultados.

```r
# Crear una fila separada para los valores F y p-value con significancia
anova_row <- data.frame(
  LForm = "F",
  `ROE_Total_sample` = paste0(
    anova_results$F.value..ROE.[1],
    " (", formatC(anova_results$p.value..ROE.[1], format = "f", digits = 4), ")"
  ),
  `ROA_Total_sample` = paste0(
    anova_results$F.value..ROA.[1],
    " (", formatC(anova_results$p.value..ROA.[1], format = "f", digits = 4), ")"
  ),
  `ROE_Spain` = paste0(
    anova_results$F.value..ROE.[2],
    " (", formatC(anova_results$p.value..ROE.[2], format = "f", digits = 4), ")"
  ),
  `ROA_Spain` = paste0(
    anova_results$F.value..ROA.[2],
    " (", formatC(anova_results$p.value..ROA.[2], format = "f", digits = 4), ")"
  ),
  `ROE_Italy` = paste0(
    anova_results$F.value..ROE.[3],
    " (", formatC(anova_results$p.value..ROE.[3], format = "f", digits = 4), ")"
  ),
  `ROA_Italy` = paste0(
    anova_results$F.value..ROA.[3],
    " (", formatC(anova_results$p.value..ROA.[3], format = "f", digits = 4), ")"
  )
)
```

Se realiza una transformación en las tablas previas para asegurar su correcta visualización. Primero, las columnas numéricas se convierten en caracteres y se reemplazan valores nulos con "0" mediante replace_na(). Luego, se combinan las tablas de promedios y valores ANOVA en una única tabla consolidada (tabla_combinada). Además, los códigos numéricos de las formas legales de las empresas se reemplazan por sus nombres correspondientes utilizando un diccionario de valores (tabla5_empresa), lo que mejora la interpretación de la información.

```r
# Convertir columnas en tabla_means a character y reemplazar NA por "0"
tabla_means <- tabla_means %>%
  mutate(across(everything(), as.character)) %>%
  mutate(across(everything(), ~ replace_na(.x, "0")))

# Convertir columnas en anova_row a character y reemplazar NA por "0"
anova_row <- anova_row %>%
  mutate(across(everything(), as.character)) %>%
  mutate(across(everything(), ~ replace_na(.x, "0")))

# Combinar ambas tablas
tabla_combinada <- bind_rows(tabla_means, anova_row)


tabla5_empresa <- c("0" = "Public limited company",
                    "1" = "Private limited company",
                    "3" = "Cooperative",
                    "4" = "Other legal forms",
```

```
                    "F" = "F")
# Reemplazar los códigos numéricos con los nombres de tipos de empresas
tabla_combinada <-tabla_combinada %>%
  mutate(LForm = tabla5_empresa[as.character(LForm)])%>%
  rename("Legal_Form" = LForm)
```

Finalmente, se genera una tabla en formato visual utilizando la librería gt(). Se establecen títulos y subtítulos con formato Markdown para mejorar la presentación del informe. Se organizan las columnas bajo tres categorías: muestra total, España e Italia, utilizando tab_spanner().Se configuran estilos de negrita para los grupos de filas y se alinean los encabezados y datos en el centro. Además, se añaden notas al pie para clarificar que los valores p aparecen entre paréntesis y que los niveles de significancia se indican con asteriscos. Finalmente, se ajustan los tamaños de fuente mediante tab_options(), asegurando que la tabla sea clara y legible.

```
# Crear la tabla con `gt()`
tabla_combinada %>%
  gt(rowname_col = "Forma Legal") %>%
  tab_header(
    title = md("**Tabla 5. Promedio de ROA y ROE por forma legal**"),
    subtitle = md("Análisis de varianza.")
  ) %>%
  cols_label(
    Legal_Form = "FORMA LEGAL",
    ROE_Total_sample = "ROE",
    ROA_Total_sample = "ROA",
    ROE_Spain = "ROE",
    ROA_Spain = "ROA",
    ROE_Italy = "ROE",
    ROA_Italy = "ROA"
  ) %>%
  tab_spanner(
    label = "Muestra total",
    columns = c(ROE_Total_sample, ROA_Total_sample)
  ) %>%
  tab_spanner(
    label = "España",
    columns = c(ROE_Spain, ROA_Spain)
  ) %>%
  tab_spanner(
    label = "Italia",
    columns = c(ROE_Italy, ROA_Italy)
  ) %>%
  tab_style(
    style = list(
      cell_text(weight = "bold")
    ),
    locations = cells_row_groups()
  ) %>%
  opt_table_lines(extent = "all") %>%
  opt_align_table_header(align = "center") %>%
  cols_align(
    align = "center",
    columns = everything()
```

Tabla 5. Promedio de ROA y ROE por forma legal

Análisis de varianza.[1,2]

	Muestra total		España		Italia	
FORMA LEGAL	ROE	ROA	ROE	ROA	ROE	ROA
Public limited company	2.67	1.62	2.17	1.39	4.69	2.5
Private limited company	5.5	3.4	3.12	3.75	6.9	3.18
Cooperative	4.46	1.64	2.91	1.39	5.05	1.74
Other legal forms	-0.99	0.73	0	0	-0.99	0.73
F	0.48 (0.6989)	6.2*** (0.0003)	0.03 (0.9709)	7.75*** (0.0005)	0.38 (0.7666)	1.6 (0.1870)

[1] *p-valor entre paréntesis.*
[2] ***, ** y * denotan un nivel de significancia por debajo del 1%, 5% y 10%, respectivamente.*
Fuente: Elaboración propia.

```
  ) %>%
  tab_footnote(
    footnote = md("*p-valor entre paréntesis.*"),
    locations = cells_title(groups = "subtitle")
  ) %>%
  tab_footnote(
    footnote = md("***, ** y * denotan un nivel de significancia
                  por debajo del 1%, 5% y 10%, respectivamente.*"),
    locations = cells_title(groups = "subtitle")
  ) %>%
  tab_source_note(
    source_note = "Fuente: Elaboración propia."
  ) %>%
  tab_options(
    column_labels.font.size = px(10), # Tamaño del texto del encabezado
    table.font.size = px(10) # Tamaño del texto en la tabla
  )%>%
  tab_style(
    style = list(
      cell_text(align = "center")
    ),
    locations = cells_title(groups = "title")
  )
```