UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

ESCOLA TÉCNICA SUPERIOR DE ENXEÑARÍA

Análise descritivo de conxunto de datos

Coautores:
Alicia Jiajun Lorenzo Lourido
Abraham Trashorras Rivas

Máster Universitario en Tecnoloxías de Análise de Datos Masivos: Big Data

Outubro 2023

Traballo para a materia de Análise Estatístico

Análise

1. Introdución

Neste análise empregamos datos do INE, en específico os datos de sociedades mercantes constituídas, clasificadas por comunidade autónoma entre os anos 2002 e 2022 dispoñibles no seguinte enlace. Modificamos previamente os datos para eliminar columnas innecesarias así como til, estes datos están dispoñibles no ficheiro *Datos_Sociedades.csv*. Unha vez importados, modificámolos ata obter a estrutura de columnas:

- **Lugar**: Comunidade ou Cidade Autónoma á cal pertencen os datos. Variable cualitativa nominal.
- **Periodo**: ano da medida. Variable cuantitativa discreta.
- **NumeroSociedades**: número de sociedades que existían. Variable cuantitativa discreta.
- **CapitalSuscrito**: capital comprometido a sociedades en miles de euros. Variable cuantitativa continua.
- **CapitalDesembolsado**: capital invertido en sociedades en miles de euros. Variable cuantitativa continua.

O noso obxectivo é analizar a evolución dos datos para Galicia entre os anos 2009 a 2022, especialmente ver o efecto do confinamento debido á pandemia do COVID-19 no ano 2020, polo que traballamos con dous subconxuntos dos datos: aqueles que pertencen a Galicia entre os anos 2009 e 2022 ambos inclusive, chamado galicia_con2020, e o mesmo subconxunto excluindo os datos do ano 2020, chamado galicia_sen2020. Isto déixanos con dous subconxuntos de 4 variables efectivas, xa que Lugar sempre terá o valor "12 Galicia", con 14 e 13 filas respectivamente.

2. Análise descritivo

Nos cadros 1, 2 e 3 podemos ver para o Número de Sociedades, o Capital Suscrito e o Capital Desembolsado respectivamente as medidas de mínimo, media, mediana, máxima, desviación típica y simetría para os dous subconxuntos.

Subconxunto	Mín	Media	Mediana	Máx	SD	Simetria
galicia_con2020	3212	3994	4032	4533	306	-0.83
galicia_sen2020	3772	4054	4040	4533	216	0.67

Cadro 1: Estudo dos valores NumeroSociedades para os dous subconxuntos

Podemos ver no Cadro 1 como o ano 2020 representa o mínimo para o Número de Sociedades cun dato tan anómalo que modifica totalmente a simetria dos datos. En contraste, estudando a desviación típica e a simetria dos Cadros 2 e 3 podemos afirmar que os datos do ano 2020 non representan un dato anómalo e tampouco é un valor mínimo. A maiores, o capital prometido ou suscrito tende a ser maior que o capital realmente invertido ou desembolsado, con maior media, mediana e máximo.

Subconxunto	Mín	Media	Mediana	Máx	SD	Simetria
galicia_con2020	140190	270746	242101	477676	96450	0.64
galicia_sen2020	140190	263524	238305	477676	96368	0.85

Cadro 2: Estudo dos valores CapitalSuscrito para os dous subconxuntos

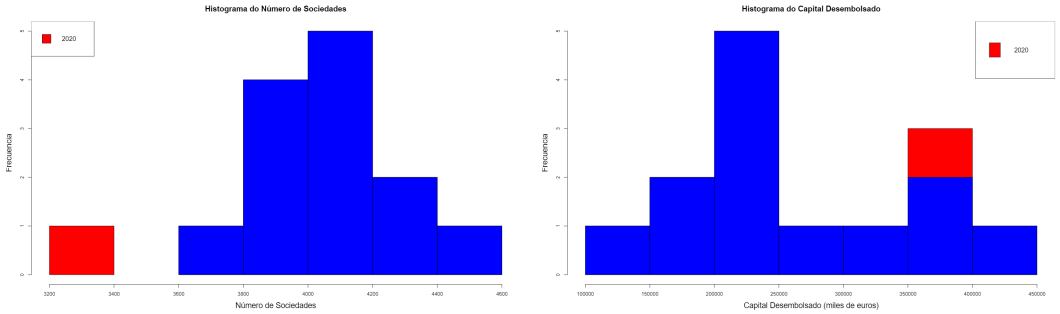
Subconxunto	Mín	Media	Mediana	Máx	SD	Simetria
galicia_con2020	139064	265348	241270	414233	87807	0.31
galicia_sen2020	139064	257720	237645	414233	86430	0.49

Cadro 3: Estudo dos valores CapitalDesembolsado para os dous subconxuntos

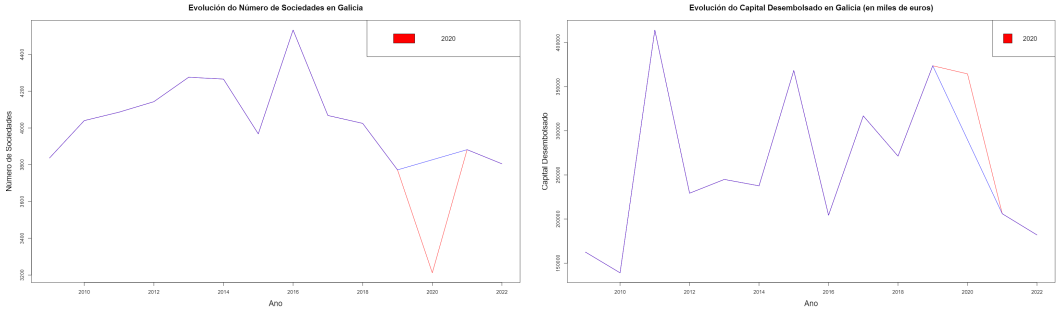
3. Análise gráfico

Tras analizar numericamente os datos, decidimos realizar o estudo gráfico na Figura 1 do Número de Sociedades e do Capital Desembolsado xa que este último e un valor real en comparación á promesa do Capital Subscrito.

Os histogramas confirman nas subfiguras (a) e (b) que mentres que o ano 2020 foi o peor entre os estudados para o Número de Sociedades, foi un ano moi positivo para o Capital Desembolto. Observando o comportamento destes dous nas Gráficas de liñas (c) e (d) podemos afirmar que no ano 2020 un gran número de sociedades pecharon a pesar dunha enorme inversión económica, mentres que no ano 2021 revertese a tendencia cunha gran creación de novas sociedades e unha caída da inversión económica.



(a) Histograma Número Sociedades (b) Histograma Capital Desembolto



(c) Gráfico de liñas Número Sociedades (d) Gráfico de liñas Capital Desenvolto

Figura 1: Visualización dos datos para galicia, anos 2009 a 2022. Ano 2020 en vermello. Imaxes dispoñibles no Anexo A

4. Inferencia

Finalmente, queremos investigar se o ano 2020 presentou unha desviación da media tan grande que modifique esta fora dun rango de confianza, sendo o confinamento un suceso altamente disrruptivo. Para isto, realizamos unha proba de duas mostras de Welch sobre os datos do Número de sociedades dos dous subconxuntos xa que vimos que para os Capitais Suscrito e Desenvolto o impacto non foi tal.

O resultado, incluído no Codigo 1 do anexo, é dunha diferencia entre medias de entre -268 e 149 sociedades os cales son uns valores de escasa dimensión en comparación as medias de 3994 e 4054 sociedades respectivamente, ademais de que é de relevancia a inclusión do 0 dentro do rango o que podería indicar que non existe diferencia.

Polo tanto, podemos dicir que aínda que anómalo o ano 2020 non presentou un valor que non puidese ser previsto nunha perdición pesimista do mercado baseándose no comportamento dos outros anos.

Anexo A

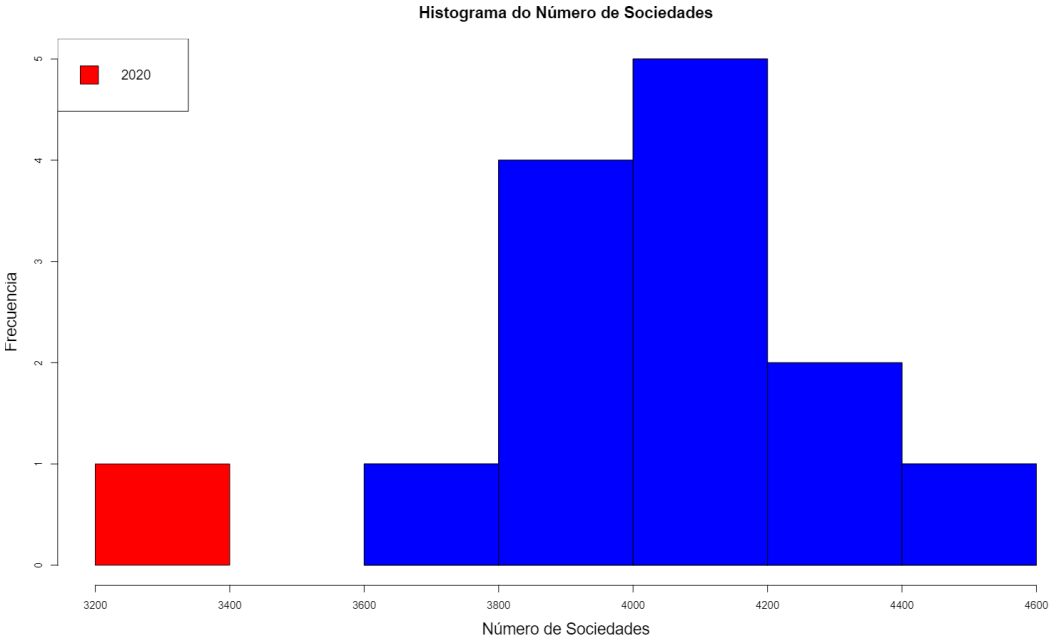


Figura 2: Histograma Número Sociedades

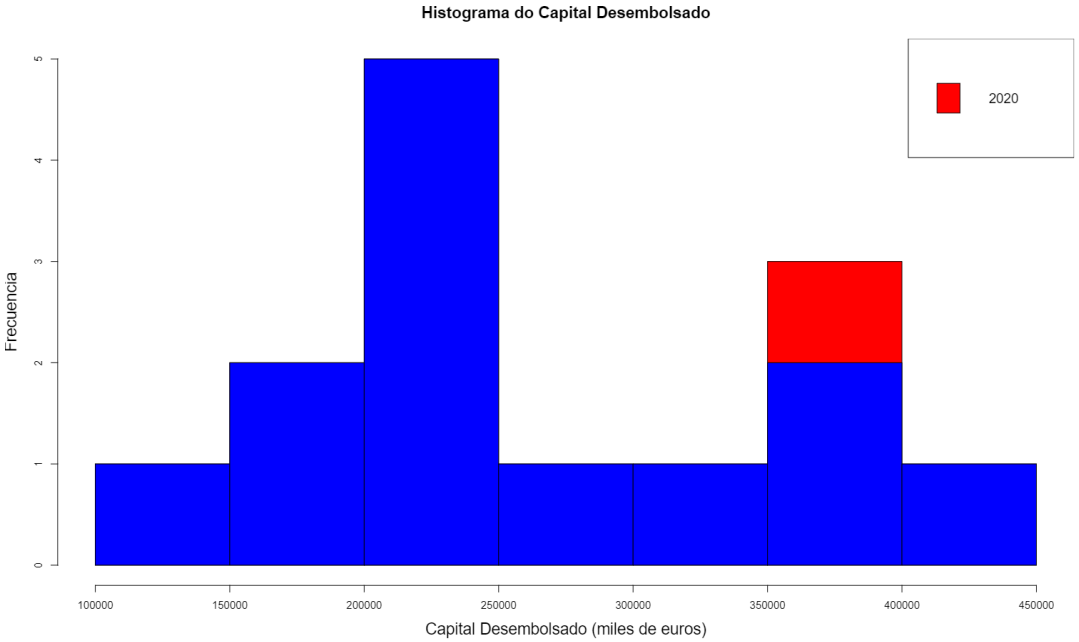


Figura 3: Histograma Capital Desembolto

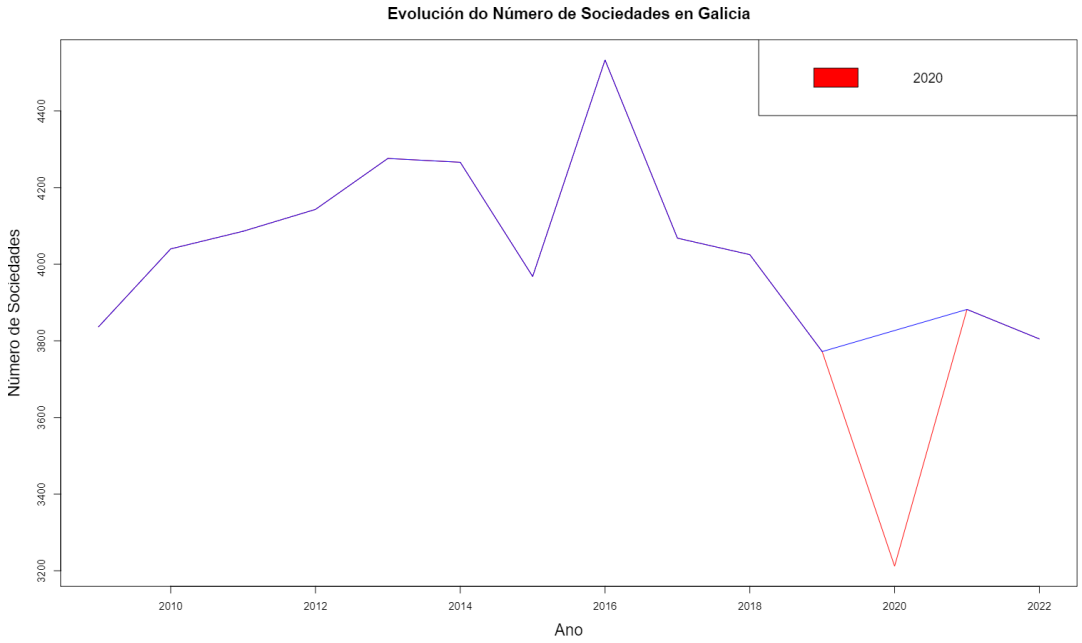


Figura 4: Gráfico de liñas Número Sociedades

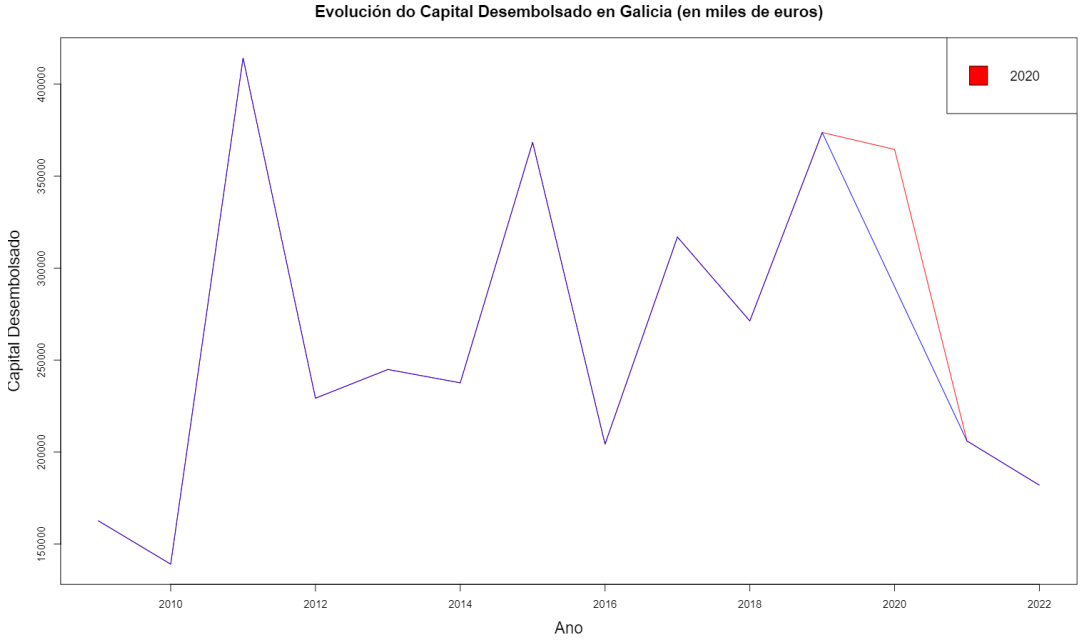


Figura 5: Gráfico de liñas Capital Desenvolto

Listing 1: Codigo e salida do test Welch

```
t.test(galicia_con2020$NumeroSociedades,
    galicia_sen2020$NumeroSociedades,
    conf.level = 0.95)

        Welch Two Sample t-test

data:  galicia_con2020$NumeroSociedades
    and galicia_sen2020$NumeroSociedades
t = -0.59304, df = 23.394, p-value = 0.5588
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -269.6891  149.4254
sample estimates:
mean of x mean of y
 3993.714  4053.846
```