

PRY_CATEDRA

Marcelo Yañez

Cátedra 1: Informe de Análisis de Datos

Definición del problema/objetivo de investigación.

Introducción:

El Servicio de Impuesto internos (SII), entrega la nómina de las personas jurídicas y las **actividades económicas** asociadas a ellas, además de sus direcciones, con esto se puede ver cómo se concentra la industria en distintos sectores, según el código que presenta.

Dicha actividad económica es una adaptación chilena del instituto nacional de estadística (INE) a la clasificación industrial internacional uniforme (CIIU), la cual clasifica las actividades económicas a nivel local, nacional e internacional.

Esta clasificación ayuda a la organización de datos relacionados con las actividades de la industria, como la agricultura, manufactura u otros los denominados Sectores Económicos.

https://www.ine.gob.cl/docs/default-source/buenas-practicas/clasificaciones/ciiu/clasificador/ciiu4-cl-2012.pdf?sfvrsn=509b707c_3

En resumen, el Código SII de actividad económica es un Código de largo 6 donde:

Cada código está representado de la siguiente forma:

- **Sección:** Letra
- **División:** 2 dígitos
- **Grupo:** 3 dígitos
- **Clase:** 4 dígitos (código final)

Sección	Division	Grupo	Clase	Subclase	Codigo SII
A	01	011	0111	01110	011101

Cada sección le corresponden todas las actividades económicas que empiezan según su división que a su vez cada división se agrupa por grupo, clase y subclase.

Sección	Glosa	Divisiones
A	Agricultura, ganadería, silvicultura y pesca	01 a 03
B	Explotación de minas y canteras	04 a 09
C	Industrias manufactureras	10 a 33
D	Suministro de electricidad, gas, vapor y aire acondicionado	35
E	Suministro de agua; evacuación de aguas residuales, gestión de desechos y descontaminación	36 a 39
F	Construcción	41 a 43
G	Comercio al por mayor y al por menor; reparación de vehículos automotores y motocicletas	45 a 47
H	Transporte y almacenamiento	49 a 53
I	Actividades de alojamiento y de servicio de comidas	55 a 56
J	Información y comunicaciones	58 a 63
K	Actividades financieras y de seguros	64 a 66
L	Actividades inmobiliarias	68
M	Actividades profesionales, científicas y técnicas	69 a 75
N	Actividades de servicios administrativos y de apoyo	77 a 82
O	Administración pública y defensa; planes de seguridad social de afiliación obligatoria	84
P	Enseñanza	85
Q	Actividades de atención de la salud humana y de asistencia social	86 a 88
R	Actividades artísticas, de entretenimiento y recreativas	90 a 93
S	Otras actividades de servicios	94 a 96
T	Actividades de los hogares como empleadores; actividades no diferenciadas de los hogares como productores de bienes y servicios para uso propio	97 a 98
U	Actividades de organizaciones y órganos extraterritoriales	99

Los datos serán recopilados desde el sitio público del Servicio de impuestos internos (SII)

https://www.sii.cl/sobre_el_sii/nominapersonasjuridicas.html.

- Nómina de actividades económicas de contribuyentes personas jurídicas

Archivo PUB_NOM_ACTECOS.txt

Listado de todos los contribuyentes personas jurídicas, junto al listado de todas las actividades económicas inscritas y vigentes a la fecha de publicación de los datos.

(Actualización: mayo 2025)

3.574.193 registros

https://www.sii.cl/estadisticas/nominas/PUB_NOM_DIRECCIONES.zip

Listado de todos los contribuyentes personas jurídicas, junto al listado de todas las direcciones registradas (Casa Matriz) vigentes y no vigentes a la fecha de publicación de los datos.

(Actualización: mayo 2025)

PUB_NOM_DOMICILIO.txt

4.682.116 Registros

Marco Teórico:

El conjunto representa empresas chilenas, proveniente del Servicio de Impuestos internos, con columnas que indican identificadores de empresas (RUT), actividades económicas (CODIGO ACTIVIDAD, “Division, Clase”), y direcciones de estas.

Análisis Exploratorio de Datos (EDA): Proceso para resumir características de los datos, visualizar distribuciones e identificar patrones .

Clasificación de Actividades Económicas: Basada en la Clasificación Industrial Internacional Uniforme (CIIU), donde Division y Clase representan sectores económicos jerárquicos.

Análisis Temporal: Estudio de tendencias a lo largo del tiempo usando las fechas para entender el crecimiento empresarial o cambios sectoriales.

Análisis Regional: Investigación de la distribución geográfica para identificar centros económicos.

Objetivo del análisis

El análisis se basa en la geografía económica y la demografía empresarial, con el objetivo de entender patrones en las actividades económicas.

El conjunto de datos CIIU (Clasificación Industrial Internacional Uniforme) para categorizar actividades económicas.

Entregables

1. **Informe en Quarto:** Este documento con código, visualizaciones e interpretaciones.
2. **Visualizaciones:** Gráficos de barras y líneas mostrando distribuciones y tendencias de actividades.
3. **Principales actividades:** métricas por actividades y región.
4. **Dashboard en Shiny** ≥ 3 Características. (Proyecto Final)

Metodología

Técnicas para el Análisis Exploratorio de Datos en R

Carga de datos

Inspección de datos

Limpieza de datos

Análisis de distribuciones

Detección de valores atípicos

Análisis de correlaciones

Transformaciones y normalización

Visualización de datos

Carga de datos

```
# Se instalan Paquetes y librerías
# Instalar si no las tienes:
# install.packages("tidyvers")
library(tibble)# Proporciona una versión mejorada de los data frames
library(ggplot2)# graficos
library(dplyr) # Herramienta para manipular datos de forma eficiente
```

Adjuntando el paquete: 'dplyr'

The following objects are masked from 'package:stats':

filter, lag

The following objects are masked from 'package:base':

intersect, setdiff, setequal, union

```
library(readr) # Permite leer y escribir archivos de datos (como CSV, TSV, TXT)
library(stringr) #Para manipular cadenas de texto de manera consistente y eficiente
library(purrr) #Facilita la programación funcional, permitiendo iterar sobre listas, vectores
library(knitr) # Para el formato de tablas.
```

Carga de datos PUB_NOM_ACTECOS.txt

```
PUB_NOM_ACTECOS <- read_delim("PUB_NOM_ACTECOS.txt",
  delim = "\t", escape_double = FALSE, trim_ws = TRUE)
```

Rows: 3574192 Columns: 7

-- Column specification -----

Delimiter: "\t"

chr (5): DV, CODIGO ACTIVIDAD, DESC. ACTIVIDAD ECONOMICA, FECHA, AFECTA A IVA

```
dbl (2): RUT, CATEGORIA TRIBUTARIA
```

```
i Use `spec()` to retrieve the full column specification for this data.  
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Guardamos el tibble

```
PUB_NOM_ACTECOS<- as_tibble(PUB_NOM_ACTECOS)
```

Carga de datos PUB_NOM_DOMICILIO.txt

```
PUB_NOM_DOMICILIO <- read_delim("PUB_NOM_DOMICILIO.txt",delim = "\t", escape_double = FALSE,
```

```
Rows: 4682114 Columns: 13
```

```
-- Column specification -----
```

```
Delimiter: "\t"
```

```
chr (11): DV, VIGENCIA, TIPO_DIRECCION, CALLE, NUMERO, BLOQUE, DEPARTAMENTO...
```

```
dbl (1): RUT
```

```
date (1): FECHA
```

```
i Use `spec()` to retrieve the full column specification for this data.  
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Guardamos el tibble

```
PUB_NOM_DOMICILIO<- as_tibble(PUB_NOM_DOMICILIO)
```

Inspección de datos

- PUB_NOM_ACTECOS

```
# vemos los primeros registros PUB_NOM_ACTECOS  
head(PUB_NOM_ACTECOS)
```

```
# A tibble: 6 x 7
```

	RUT	DV	`CODIGO ACTIVIDAD`	DESC. ACTIVIDAD ECONO~1	FECHA	`AFECTA A IVA`
	<dbl>	<chr>	<chr>	<chr>	<chr>	<chr>
1	50000066	K	643000	FONDOS Y SOCIEDADES DE~	01-0~	S
2	50000100	3	879000	OTRAS ACTIVIDADES DE A~	01-0~	S
3	50000430	4	472101	VENTA AL POR MENOR DE ~	01-0~	S

```

4 50000510 6      492300      TRANSPORTE DE CARGA PO~ 01-0~ S
5 50000710 9      251100      FABRICACION DE PRODUCT~ 01-0~ S
6 50001570 5      561000      ACTIVIDADES DE RESTAUR~ 01-0~ S
# i abbreviated name: 1: `DESC. ACTIVIDAD ECONOMICA`
# i 1 more variable: `CATEGORIA TRIBUTARIA` <dbl>

```

```

# Resumen PUB_NOM_ACTECOS
str(PUB_NOM_ACTECOS)

```

```

tibble [3,574,192 x 7] (S3: tbl_df/tbl/data.frame)
 $ RUT                : num [1:3574192] 5e+07 5e+07 5e+07 5e+07 5e+07 ...
 $ DV                 : chr [1:3574192] "K" "3" "4" "6" ...
 $ CODIGO ACTIVIDAD   : chr [1:3574192] "643000" "879000" "472101" "492300" ...
 $ DESC. ACTIVIDAD ECONOMICA: chr [1:3574192] "FONDOS Y SOCIEDADES DE INVERSION Y ENTIDADES I
 $ FECHA              : chr [1:3574192] "01-01-1993" "01-01-1993" "01-01-1993" "01-01-
 $ AFECTA A IVA       : chr [1:3574192] "S" "S" "S" "S" ...
 $ CATEGORIA TRIBUTARIA : num [1:3574192] 1 1 1 1 1 1 1 1 1 1 ...

```

Estructura PUB_NOM_ACTECOS

RUT (chr): (cualitativa, nominal): Identificador único del contribuyente.

DV (chr): (cualitativa, nominal): Dígito verificador del RUT.

CODIGO ACTIVIDAD (chr): (cualitativa, nominal): Código de la actividad económica.

DESC. ACTIVIDAD ECONOMICA (chr): (cualitativa, nominal): Descripción de la actividad.

FECHA (chr): (cuantitativa, continua): Fecha de registro.

AFECTA A IVA (chr): (cualitativa, nominal): Indica si la actividad afecta al IVA (“S” o “N”).

CATEGORIA TRIBUTARIA (chr): (cualitativa, ordinal): Categoría tributaria (1, 2,).

- PUB_NOM_DOMICILIO

```

head(PUB_NOM_DOMICILIO)

```

```

# A tibble: 6 x 13
      RUT DV  VIGENCIA FECHA      TIPO_DIRECCION CALLE      NUMERO BLOQUE
  <dbl> <chr> <chr>    <date>    <chr>          <chr>      <chr> <chr>
1 47003012 7    S      2010-05-07 DOMICILIO      AV. BOSQUE NO~ 0177 <NA>
2 47003030 5    N      2014-06-12 DOMICILIO      ENRIQUE FOSTER 118 <NA>

```

```

3 47003030 5      N      2014-01-24 DOMICILIO      ISIDORA GOYEN~ 3520      <NA>
4 47003030 5      S      2014-07-03 DOMICILIO      SANTA LUCIA    268      <NA>
5 47014441 6      S      2022-05-03 DOMICILIO      ISIDORA GOYEN~ 3520      <NA>
6 47014804 7      S      2013-03-14 DOMICILIO      AVDA. APOQUIN~ 3457      <NA>
# i 5 more variables: DEPARTAMENTO <chr>, VILLA_POBLACION <chr>, CIUDAD <chr>,
#   COMUNA <chr>, REGION <chr>

```

```

# Resumen PUB_NOM_DOMICILIO
str(PUB_NOM_DOMICILIO)

```

```

tibble [4,682,114 x 13] (S3: tbl_df/tbl/data.frame)
 $ RUT          : num [1:4682114] 4.7e+07 4.7e+07 4.7e+07 4.7e+07 4.7e+07 ...
 $ DV           : chr [1:4682114] "7" "5" "5" "5" ...
 $ VIGENCIA     : chr [1:4682114] "S" "N" "N" "S" ...
 $ FECHA        : Date[1:4682114], format: "2010-05-07" "2014-06-12" ...
 $ TIPO_DIRECCION : chr [1:4682114] "DOMICILIO" "DOMICILIO" "DOMICILIO" "DOMICILIO" ...
 $ CALLE        : chr [1:4682114] "AV. BOSQUE NORTE" "ENRIQUE FOSTER" "ISIDORA GOYENECHEA" ...
 $ NUMERO       : chr [1:4682114] "0177" "118" "3520" "268" ...
 $ BLOQUE       : chr [1:4682114] NA NA NA NA ...
 $ DEPARTAMENTO  : chr [1:4682114] "P.3" "2004" "PISO3" "401" ...
 $ VILLA_POBLACION: chr [1:4682114] NA NA NA NA ...
 $ CIUDAD       : chr [1:4682114] "SANTIAGO" "SANTIAGO" "SANTIAGO" "SANTIAGO" ...
 $ COMUNA       : chr [1:4682114] "LAS CONDES" "LAS CONDES" "LAS CONDES" "SANTIAGO" ...
 $ REGION      : chr [1:4682114] "XIII REGION METROPOLITANA" "XIII REGION METROPOLITANA" ...

```

Estructura PUB_NOM_DOMICILIO

RUT (chr): (cualitativa, nominal): Identificador único del contribuyente.

DV (chr): (cualitativa, nominal): Dígito verificador del RUT.

VIGENCIA (chr): (cualitativa, nominal): Indica si el domicilio está vigente (“S” o “N”).

FECHA (Date): (cuantitativa, continua): Fecha de registro.

TIPO_DIRECCION (chr): (cualitativa, nominal): Tipo de dirección (“MATRIZ= DOMICILIO”).

CALLE (chr): (cualitativa, nominal): Nombre de la calle.

NUMERO (chr): (cualitativa, nominal): Número de la dirección.

BLOQUE (chr): (cualitativa, nominal): Bloque del domicilio

DEPARTAMENTO (chr): (cualitativa, nominal): Número de departamento

VILLA_POBLACION (chr): (cualitativa, nominal)

CIUDAD (chr): (cualitativa, nominal): Ciudad del domicilio.

COMUNA (chr): (cualitativa, nominal): Comuna del domicilio.

REGION (chr): (cualitativa, nominal): Región del domicilio

Limpieza de datos Transformaciones y normalización

- PUB_NOM_ACTECOS

FECHA viene en formato (chr), se transformará en Date, para el mejor manejo de fechas.

Se guardara en nuevo frame PUB_NOM_ACTECOS2

```
# Transformamos la fecha chr en date y la guardamos en nuevo frame actualizado PUB_NOM_ACTECOS2
```

```
PUB_NOM_ACTECOS2 <- PUB_NOM_ACTECOS %>%  
  mutate(FECHA = as.Date(FECHA, format = "%d-%m-%Y"))
```

```
# Resumen PUB_NOM_ACTECOS2  
str(PUB_NOM_ACTECOS2)
```

```
tibble [3,574,192 x 7] (S3: tbl_df/tbl/data.frame)  
  $ RUT                : num [1:3574192] 5e+07 5e+07 5e+07 5e+07 5e+07 ...  
  $ DV                 : chr [1:3574192] "K" "3" "4" "6" ...  
  $ CODIGO ACTIVIDAD   : chr [1:3574192] "643000" "879000" "472101" "492300" ...  
  $ DESC. ACTIVIDAD ECONOMICA: chr [1:3574192] "FONDOS Y SOCIEDADES DE INVERSION Y ENTIDADES I  
  $ FECHA              : Date[1:3574192], format: "1993-01-01" "1993-01-01" ...  
  $ AFECTA A IVA       : chr [1:3574192] "S" "S" "S" "S" ...  
  $ CATEGORIA TRIBUTARIA : num [1:3574192] 1 1 1 1 1 1 1 1 1 1 ...
```

```
summary(PUB_NOM_ACTECOS2)
```

RUT	DV	CODIGO ACTIVIDAD
Min. :50000066	Length:3574192	Length:3574192
1st Qu.:76534365	Class :character	Class :character
Median :77121556	Mode :character	Mode :character
Mean :76955060		
3rd Qu.:77673652		
Max. :99599860		

DESC. ACTIVIDAD ECONOMICA	FECHA	AFECTA A IVA
---------------------------	-------	--------------

Length:3574192	Min. :1196-04-19	Length:3574192
Class :character	1st Qu.:2015-04-21	Class :character
Mode :character	Median :2020-04-06	Mode :character
	Mean :2017-11-26	
	3rd Qu.:2022-09-28	
	Max. :2025-05-08	
	NA's :113	

CATEGORIA TRIBUTARIA

Min. :1.000
1st Qu.:1.000
Median :1.000
Mean :1.018
3rd Qu.:1.000
Max. :2.000

En PUB_NOM_ACTECOS2 visualizamos FECHA, hay variables fuera de rango “< a 1993-01-01”

- Cambiamos estas variables por 1993-01-01 fecha por defecto del archivo
- Por otra parte, vemos 133 datos faltantes “NA’s” se ingresará fecha por defecto mencionada.
- Para el análisis final se recopilaran solo los campos , RUT, ACTIVIDAD ECONOMICA y FECHA
- El código de actividad económica lo transformamos en Factor
- Cambiamos el nombre del campo fecha por FEC_ACT_ECO, que es más representativo

```
# Preprocesar FECHA en PUB_NOM_ACTECOS2 SE COMBIERTE CODIGO ACTIVIDAD EN FACTOR y SE DEJAN
PUB_NOM_ACTECOS2 <- PUB_NOM_ACTECOS2 %>%
  mutate(`CODIGO ACTIVIDAD` = as.factor(`CODIGO ACTIVIDAD`))%>%
  mutate(FECHA = if_else(is.na(FECHA) | FECHA < as.Date("1993-01-01"), as.Date("1993-01-01"), FECHA))%>%
  select(RUT, 'CODIGO ACTIVIDAD', FEC_ACT_ECO)
```

Subdividimos el CODIGO ACTIVIDAD y creamos los campos División y Clase, necesarios para nuestro análisis.

```
PUB_NOM_ACTECOS2<-PUB_NOM_ACTECOS2%>% mutate(
  Division = substr(as.character(PUB_NOM_ACTECOS2$CODIGO_ACTIVIDAD), 1, 3),
  Clase = substr(as.character(PUB_NOM_ACTECOS2$CODIGO_ACTIVIDAD), 4, 5))%>%
  head(PUB_NOM_ACTECOS2)
```

```
# A tibble: 6 x 5
      RUT `CODIGO ACTIVIDAD` FEC_ACT_ECO Division Clase
    <dbl> <fct>             <date>      <chr>    <chr>
1 50000066 643000          1993-01-01  64      6430
2 50000100 879000          1993-01-01  87      8790
3 50000430 472101          1993-01-01  47      4721
4 50000510 492300          1993-01-01  49      4923
5 50000710 251100          1993-01-01  25      2511
6 50001570 561000          1993-01-01  56      5610
```

Division y Clase los convertimos en factor

```
PUB_NOM_ACTECOS2<-PUB_NOM_ACTECOS2 %>% mutate(Division = as.factor(Division), Clase = as.factor(Clase))
```

- PUB_NOM_DOMICILIO

```
head(PUB_NOM_DOMICILIO)
```

```
# A tibble: 6 x 13
      RUT DV  VIGENCIA FECHA      TIPO_DIRECCION CALLE      NUMERO BLOQUE
    <dbl> <chr> <chr>    <date>      <chr>          <chr>      <chr> <chr>
1 47003012 7    S        2010-05-07 DOMICILIO      AV. BOSQUE NO~ 0177  <NA>
2 47003030 5    N        2014-06-12 DOMICILIO      ENRIQUE FOSTER 118   <NA>
3 47003030 5    N        2014-01-24 DOMICILIO      ISIDORA GOYEN~ 3520  <NA>
4 47003030 5    S        2014-07-03 DOMICILIO      SANTA LUCIA    268   <NA>
5 47014441 6    S        2022-05-03 DOMICILIO      ISIDORA GOYEN~ 3520  <NA>
6 47014804 7    S        2013-03-14 DOMICILIO      AVDA. APOQUIN~ 3457  <NA>
# i 5 more variables: DEPARTAMENTO <chr>, VILLA_POBLACION <chr>, CIUDAD <chr>,
# COMUNA <chr>, REGION <chr>
```

```
str(PUB_NOM_DOMICILIO)
```

```
tibble [4,682,114 x 13] (S3: tbl_df/tbl/data.frame)
 $ RUT          : num [1:4682114] 4.7e+07 4.7e+07 4.7e+07 4.7e+07 4.7e+07 ...
 $ DV           : chr [1:4682114] "7" "5" "5" "5" ...
 $ VIGENCIA     : chr [1:4682114] "S" "N" "N" "S" ...
 $ FECHA        : Date[1:4682114], format: "2010-05-07" "2014-06-12" ...
 $ TIPO_DIRECCION : chr [1:4682114] "DOMICILIO" "DOMICILIO" "DOMICILIO" "DOMICILIO" ...
 $ CALLE        : chr [1:4682114] "AV. BOSQUE NORTE" "ENRIQUE FOSTER" "ISIDORA GOYENECHEA" ...
 $ NUMERO       : chr [1:4682114] "0177" "118" "3520" "268" ...
 $ BLOQUE       : chr [1:4682114] NA NA NA NA ...
```

```
$ DEPARTAMENTO : chr [1:4682114] "P.3" "2004" "PISO3" "401" ...
$ VILLA_POBLACION: chr [1:4682114] NA NA NA NA ...
$ CIUDAD : chr [1:4682114] "SANTIAGO" "SANTIAGO" "SANTIAGO" "SANTIAGO" ...
$ COMUNA : chr [1:4682114] "LAS CONDES" "LAS CONDES" "LAS CONDES" "SANTIAGO" ...
$ REGION : chr [1:4682114] "XIII REGION METROPOLITANA" "XIII REGION METROPOLITANA"
```

```
summary(PUB_NOM_DOMICILIO)
```

RUT	DV	VIGENCIA	FECHA
Min. :47003012	Length:4682114	Length:4682114	Min. :1993-01-01
1st Qu.:76268130	Class :character	Class :character	1st Qu.:2011-06-29
Median :76960938	Mode :character	Mode :character	Median :2018-03-26
Mean :75551346			Mean :2015-05-06
3rd Qu.:77654652			3rd Qu.:2021-12-30
Max. :99599860			Max. :2025-05-08

TIPO_DIRECCION	CALLE	NUMERO	BLOQUE
Length:4682114	Length:4682114	Length:4682114	Length:4682114
Class :character	Class :character	Class :character	Class :character
Mode :character	Mode :character	Mode :character	Mode :character

DEPARTAMENTO	VILLA_POBLACION	CIUDAD	COMUNA
Length:4682114	Length:4682114	Length:4682114	Length:4682114
Class :character	Class :character	Class :character	Class :character
Mode :character	Mode :character	Mode :character	Mode :character

```
REGION
Length:4682114
Class :character
Mode :character
```

Vemos datos faltantes

```
colSums(is.na(PUB_NOM_DOMICILIO)) # Verifica cuántos NA hay por columna.
```

RUT	DV	VIGENCIA	FECHA	TIPO_DIRECCION
0	0	0	0	0
CALLE	NUMERO	BLOQUE	DEPARTAMENTO	VILLA_POBLACION
14	977448	4432886	3344147	3444161
CIUDAD	COMUNA	REGION		
1327307	0	0		

Para nuestro análisis de la data de PUB_NOM_DOMICILIO: necesitaremos los campos RUT, FECHA y REGION con estado de VIGENCIA= S “están vigentes”

```
PUB_NOM_DOMICILIO2 <- PUB_NOM_DOMICILIO %>%
  filter(VIGENCIA == "S") %>%
  select(RUT, FECHA, REGION)
```

```
str(PUB_NOM_DOMICILIO2)
```

```
tibble [3,031,724 x 3] (S3: tbl_df/tbl/data.frame)
 $ RUT      : num [1:3031724] 4.7e+07 4.7e+07 4.7e+07 4.7e+07 4.7e+07 ...
 $ FECHA    : Date[1:3031724], format: "2010-05-07" "2014-07-03" ...
 $ REGION   : chr [1:3031724] "XIII REGION METROPOLITANA" "XIII REGION METROPOLITANA" "XIII REG
```

- Cambiamos el nombre del campo fecha por FEC_DOM, que es más representativo
- Eliminamos REGION en los datos de REGION ya que se representa por el nombre de la columna y facilita su visualización.

```
PUB_NOM_DOMICILIO2 <- PUB_NOM_DOMICILIO2 %>% rename(FEC_DOM = FECHA)%>%
  mutate(REGION = str_replace(REGION, " REGION ", " "))
```

```
head(PUB_NOM_DOMICILIO2)
```

```
# A tibble: 6 x 3
   RUT FEC_DOM REGION
  <dbl> <date>   <chr>
1 47003012 2010-05-07 XIII METROPOLITANA
2 47003030 2014-07-03 XIII METROPOLITANA
3 47014441 2022-05-03 XIII METROPOLITANA
4 47014804 2013-03-14 XIII METROPOLITANA
5 47026836 2025-04-14 XIII METROPOLITANA
6 50000000 2006-10-13 XIII METROPOLITANA
```

Se cargan los niveles de las actividades economicas

```
Nvl_Act_eco <- read_delim("Nivel_Act_eco.txt",  
  delim = "\t", escape_double = FALSE, trim_ws = TRUE)
```

Rows: 419 Columns: 6

-- Column specification -----

Delimiter: "\t"

chr (6): Seccion, Gls_Seccion, Grupo, Gls_Grupo, Clase, Gls_Clase

i Use `spec()` to retrieve the full column specification for this data.

i Specify the column types or set `show_col_types = FALSE` to quiet this message.

```
head(Nvl_Act_eco)
```

A tibble: 6 x 6

	Seccion	Gls_Seccion	Grupo	Gls_Grupo	Clase	Gls_Clase
	<chr>	<chr>	<chr>	<chr>	<chr>	<chr>
1	A	Agricultura, ganadería, silvicultura	~ 011	Cultivo	~ 0111	Cultivo ~
2	A	Agricultura, ganadería, silvicultura	~ 011	Cultivo	~ 0112	Cultivo ~
3	A	Agricultura, ganadería, silvicultura	~ 011	Cultivo	~ 0113	Cultivo ~
4	A	Agricultura, ganadería, silvicultura	~ 011	Cultivo	~ 0114	Cultivo ~
5	A	Agricultura, ganadería, silvicultura	~ 011	Cultivo	~ 0115	Cultivo ~
6	A	Agricultura, ganadería, silvicultura	~ 011	Cultivo	~ 0116	Cultivo ~

Se unen PUB_NOM_ACTECOS2 y PUB_NOM_DOMICILIO2, en nuevo frame para el análisis final

```
rm(Data_Studio)
```

Warning in rm(Data_Studio): objeto 'Data_Studio' no encontrado

```
Data_Studio <- merge(PUB_NOM_ACTECOS2, PUB_NOM_DOMICILIO2, by = "RUT")
```

```
head(Data_Studio)
```

	RUT	CODIGO	ACTIVIDAD	FEC_ACT_ECO	Division	Clase	FEC_DOM
1	50000066	643000	1993-01-01	64	6430	2011-10-18	
2	50000100	879000	1993-01-01	87	8790	2006-01-30	

3	50000430	472101	1993-01-01	47	4721	1993-01-01
4	50000510	492300	1993-01-01	49	4923	1993-01-01
5	50000710	251100	1993-01-01	25	2511	1997-04-17
6	50001570	561000	1993-01-01	56	5610	1993-01-01

	REGION
1	V VALPARAISO
2	XIII METROPOLITANA
3	IV COQUIMBO
4	IV COQUIMBO
5	IV COQUIMBO
6	XIII METROPOLITANA

```
summary(Data_Studio)
```

RUT	CODIGO ACTIVIDAD	FEC_ACT_ECO	Division
Min. :50000066	702000 : 112342	Min. :1993-01-01	47 : 714834
1st Qu.:76534431	492300 : 106469	1st Qu.:2015-04-21	43 : 256776
Median :77121569	643000 : 101647	Median :2020-04-06	46 : 251867
Mean :76955565	433000 : 92370	Mean :2017-11-27	49 : 164436
3rd Qu.:77673603	681012 : 91759	3rd Qu.:2022-09-28	68 : 152490
Max. :99599860	829900 : 88558	Max. :2025-05-08	56 : 124119
	(Other):2981027		(Other):1909650

Clase	FEC_DOM	REGION
4721 : 142446	Min. :1993-01-01	Length:3574172
6810 : 128384	1st Qu.:2016-04-28	Class :character
4773 : 127612	Median :2020-07-23	Mode :character
7020 : 112342	Mean :2018-09-08	
4923 : 106469	3rd Qu.:2023-01-25	
6430 : 101647	Max. :2025-05-08	
(Other):2855272		

Estructura y tipos de datos Data_Studio

```
str(Data_Studio)
```

```
'data.frame': 3574172 obs. of 7 variables:
 $ RUT : num 5e+07 5e+07 5e+07 5e+07 5e+07 ...
 $ CODIGO ACTIVIDAD: Factor w/ 672 levels "011101","011102",...: 489 624 351 417 191 448 448 ...
 $ FEC_ACT_ECO : Date, format: "1993-01-01" "1993-01-01" ...
 $ Division : Factor w/ 88 levels "01","02","03",...: 58 78 44 45 25 51 51 84 42 58 ...
 $ Clase : Factor w/ 416 levels "0111","0112",...: 308 385 232 259 123 278 278 403 ...
```

```
$ FEC_DOM      : Date, format: "2011-10-18" "2006-01-30" ...
$ REGION       : chr  "V VALPARAISO" "XIII METROPOLITANA" "IV COQUIMBO" "IV COQUIMBO" ..
```

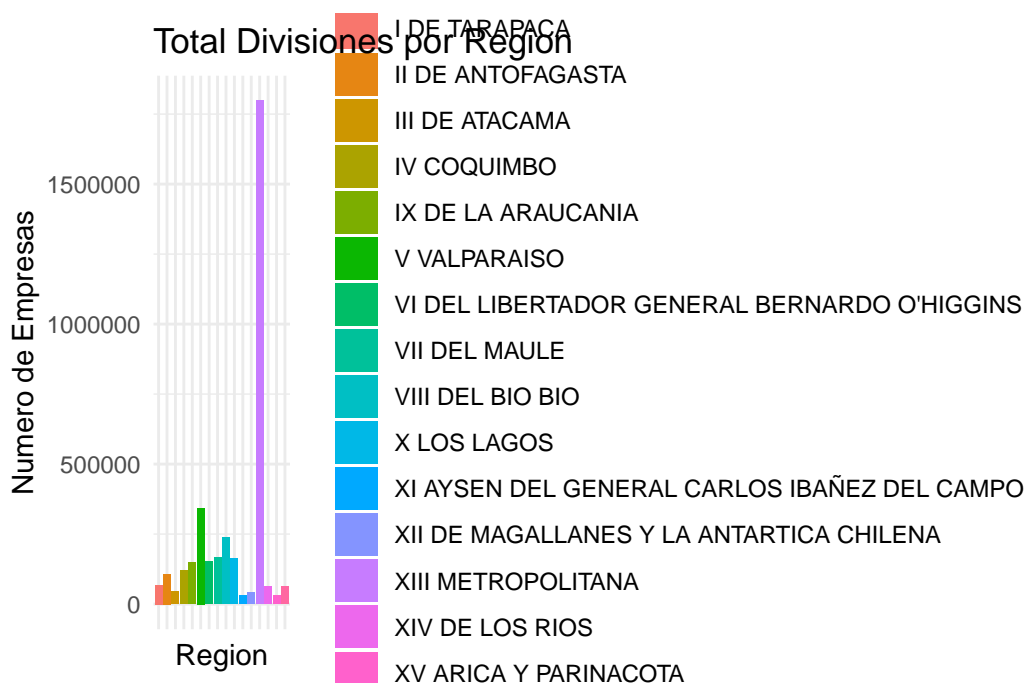
Eliminamos los registros con datos en REGION = Sin Información

```
Data_Studio <- subset(Data_Studio, REGION != "Sin Información")
```

VisualizacionesMostrar cantidad de divisiones por Region

```
data_REG_Div <- Data_Studio |>
  dplyr::group_by(REGION, Division) |>
  dplyr::summarise(n = n(), .groups = 'drop')

ggplot(data = data_REG_Div, aes(x = REGION, y = n, fill = REGION)) +
  geom_bar(stat = "identity") + labs(title = "Total Divisiones por Region", x = "Region",
  theme(axis.text.x = element_blank())
```

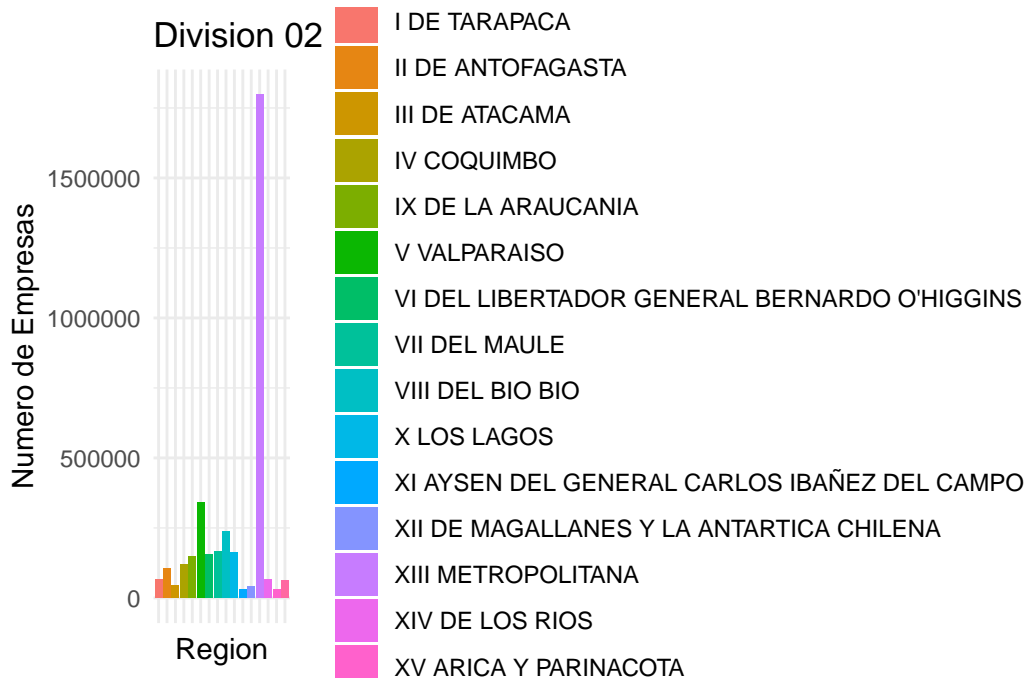


Mostrar la division = 02 por cada Region

```
data_REG_Div_02 <- Data_Studio |>
  dplyr::group_by(REGION, Division) |>
  dplyr::summarise(n = n(), .groups = 'drop') |>
  dplyr::filter(Division == "02")

ggplot(data = data_REG_Div, aes(x = REGION, y = n, fill = REGION)) +
  geom_bar(stat = "identity") + labs(title = "Division 02", x = "Region",
  theme(axis.text.x = element_blank())
```

y = "Numero de

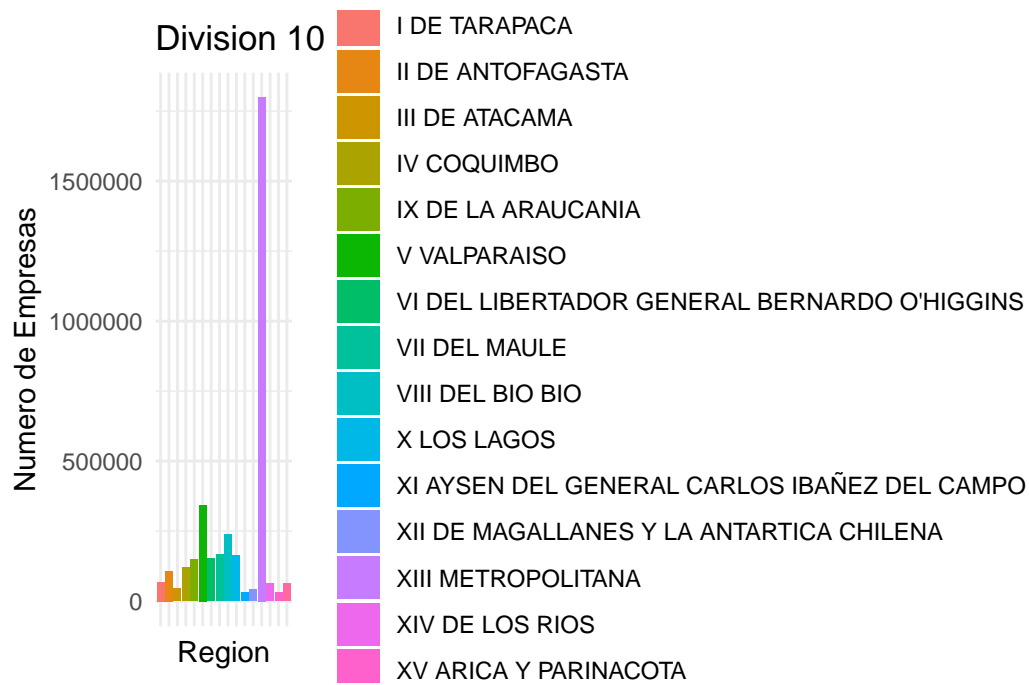


Division 10

```
data_REG_Div_10 <- Data_Studio |>
  dplyr::group_by(REGION, Division) |>
  dplyr::summarise(n = n(), .groups = 'drop') |>
  dplyr::filter(Division == "10")

ggplot(data = data_REG_Div, aes(x = REGION, y = n, fill = REGION)) +
  geom_bar(stat = "identity") + labs(title = "Division 10", x = "Region",
  theme(axis.text.x = element_blank())
```

y = "Numero de



Según la gráfica podríamos inferir que la gran concentración de empresas está en la XIII Región, independiente de su división.