Proyecto final Ciencia de Datos con R: Micro, pequeñas y medianas empresas en el Uruguay.

Gonzalo Ruiz gonzaaruizz22@gmail.com

Indice

Introducción

Datos, datos y más datos

Para el siguiente análisis utilizaremos datos de múltiple fuentes que nos ayudarán a explicar el fenómeno de las empresas de distinto tamaño en Uruguay a través de variables indicadoras de la situación económica del país y de variables relacionadas a las empresas mencionadas.

Las fuentes utilizadas fueron extraídas de:

- Agencia Nacional de Desarrollo (ANDE) (2024)
- Informe IDERE-UY (Rodríguez Miranda et al. (2024))
- Observatorio Territorio Uruguay – OPP (Observatorio Territorio Uruguay – Oficina de Planeamiento y Presupuesto (OPP) (2018))

Descripción de variables

Variable	Descripción
anio	Año de referencia de la observación.
departamento	Nombre del departamento (en Uruguay) donde se registra la información.
tamanio	Clasificación del tamaño de las empresas (ej. micro, pequeña, mediana, etc.).
sector	Sector económico al que pertenece la empresa (ej. industria, servicios, etc.).
n_empresas	Número total de empresas registradas.
n_nacimientos	Número total de nuevas empresas creadas.
n_muertes	Número de empresas que dejaron de operar.
alfabetismo	Tasa de alfabetismo.
accesos_a_estudios_terciarios	Población entre 25 y 65 años que accede a estudios terciarios(%).
anios_de_educacion_promedio	Promedio de años de educación de la población de 25 años y más.
promocion_educacion_media_cb	Porcentaje de promoción en ciclo básico de educación media pública.
pobreza	Personas en hogares en situación de pobreza (%).
informalidad	Informalidad de los ocupados (%).
desempleo_en_jovenes	Cociente entre tasa de desempleo de jóvenes (14-29 años) y tasa general.
gini	Coeficiente de Gini (ingreso de los hogares).
acceso_a_internet	Hogares con conexión a internet (%).
ingresos_de_los_hogares	Relación entre el ingreso per cápita de los hogares del departamento y el valor para el país.
tasa_de_desempleo	Tasa de desempleo.
porcentaje_de_personal_presupuestado_en_el_total_de_personal	Porcentaje de personal presupuestado en el total de funcionarios de los gobiernos departamentales.
part_act_econ	Participación porcentual del departamento en la actividad económica país.

Ánalisis exploratorio

Al trabajar con datos es natural hacerse preguntas sobre la composición y características de los mismos, es en éste siguiente análisis que intentaremos dar respuestas de manera visual y con medidas de resúmenes numéricas para familiarizarnos con nuestra base de datos.

- ¿Cuál es la dimensión de nuestro Dataset?

Tenemos un Dataset con 17775 observaciones y 20 variables.

- ¿Existen duplicados?

Tenemos 17775 observaciones y si descartamos los duplicados nos queda en: 17775.

- ¿Cómo está armado nuestro DataFrame? ¿qué variables contiene? Y ejemplos de observaciones.
  - Esquema

```
Rows: 17,775
Columns: 20
$ anio                                                  <chr> "2008", "2008",~
$ departamento                                          <fct> artigas, artiga~
$ tamanio                                               <ord> micro, micro, m~
$ sector                                                <fct> a, b, c, d, e, ~
$ n_empresas                                            <chr> "44", "27", "11~
$ n_nacimientos                                         <chr> "sin datos", "s~
$ n_muertes                                             <chr> "8", "6", "21",~
$ alfabetismo                                           <dbl> 0.7969807, 0.79~
$ accesos_a_estudios_terciarios                         <dbl> 0.1006619, 0.10~
$ anios_de_educacion_promedio                           <dbl> 0.2030340, 0.20~
$ promocion_educacion_media_cb                          <dbl> 0.5909681, 0.59~
$ pobreza                                               <dbl> 0.2828235, 0.28~
$ informalidad                                          <dbl> 0.1945994, 0.19~
$ desempleo_en_jovenes                                  <dbl> 0.7121335, 0.71~
$ gini                                                  <dbl> 0.4286669, 0.42~
$ acceso_a_internet                                     <dbl> 0.0682450, 0.06~
$ ingresos_de_los_hogares                               <dbl> 0.1288986, 0.12~
$ tasa_de_desempleo                                     <dbl> 0.5135037, 0.51~
$ porcentaje_de_personal_presupuestado_en_la_intendencia <dbl> 0.4633301, 0.46~
$ part_act_econ                                         <dbl> 1.5, 1.5, 1.5, ~
```

```
-   Primeras 6 filas
```

```
# A tibble: 6 x 20
  anio  departamento tamanio sector n_empresas n_nacimientos n_muertes
  <chr> <fct>        <ord>   <fct>  <chr>      <chr>         <chr>
1 2008  artigas      micro   a      44         sin datos     8
2 2008  artigas      micro   b      27         sin datos     6
3 2008  artigas      micro   c      114        sin datos     21
4 2008  artigas      micro   d      1          sin datos     1
5 2008  artigas      micro   e      13         sin datos     4
6 2008  artigas      micro   f      28         sin datos     6
# i 13 more variables: alfabetismo <dbl>, accesos_a_estudios_terciarios <dbl>,
#   anios_de_educacion_promedio <dbl>, promocion_educacion_media_cb <dbl>,
#   pobreza <dbl>, informalidad <dbl>, desempleo_en_jovenes <dbl>, gini <dbl>,
#   acceso_a_internet <dbl>, ingresos_de_los_hogares <dbl>,
#   tasa_de_desempleo <dbl>,
#   porcentaje_de_personal_presupuestado_en_la_intendencia <dbl>,
#   part_act_econ <dbl>
```

- Últimas 6 filas

```
# A tibble: 6 x 20
  anio  departamento   tamanio sector n_empresas n_nacimientos n_muertes
  <chr> <fct>          <ord>   <fct>  <chr>      <chr>         <chr>
1 2021  treinta y tres grandes d      sin datos  <NA>          <NA>
2 2021  treinta y tres grandes g      1          <NA>          <NA>
3 2021  treinta y tres grandes h      sin datos  <NA>          <NA>
4 2021  treinta y tres grandes k      sin datos  <NA>          <NA>
5 2021  treinta y tres grandes l      sin datos  <NA>          <NA>
6 2021  treinta y tres grandes q      1          <NA>          <NA>
# i 13 more variables: alfabetismo <dbl>, accesos_a_estudios_terciarios <dbl>,
#   anios_de_educacion_promedio <dbl>, promocion_educacion_media_cb <dbl>,
#   pobreza <dbl>, informalidad <dbl>, desempleo_en_jovenes <dbl>, gini <dbl>,
#   acceso_a_internet <dbl>, ingresos_de_los_hogares <dbl>,
#   tasa_de_desempleo <dbl>,
#   porcentaje_de_personal_presupuestado_en_la_intendencia <dbl>,
#   part_act_econ <dbl>
```

- 6 filas aleatorias

```
# A tibble: 6 x 20
  anio  departamento   tamanio  sector n_empresas n_nacimientos n_muertes
  <chr> <fct>          <ord>    <fct>  <chr>      <chr>         <chr>
1 2022  colonia        micro    r      145        28            16
```

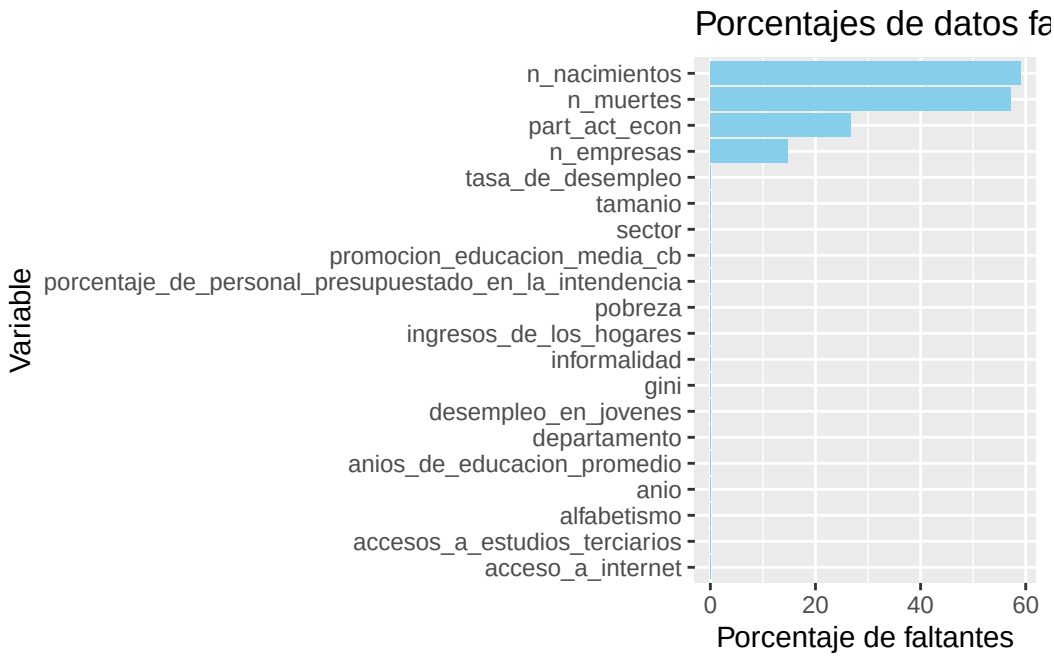
```
2 2018  soriano        grandes  p      sin datos  sin datos     <NA>
3 2022  tacuarembo     medianas j      sin datos  <NA>          <NA>
4 2010  treinta y tres pequeñas j      10         sin datos     sin datos
5 2008  lavalleja      pequeñas c      37         sin datos     2
6 2014  rivera         grandes  c      sin datos  <NA>          <NA>
# i 13 more variables: alfabetismo <dbl>, accesos_a_estudios_terciarios <dbl>,
#   anios_de_educacion_promedio <dbl>, promocion_educacion_media_cb <dbl>,
#   pobreza <dbl>, informalidad <dbl>, desempleo_en_jovenes <dbl>, gini <dbl>,
#   acceso_a_internet <dbl>, ingresos_de_los_hogares <dbl>,
#   tasa_de_desempleo <dbl>,
#   porcentaje_de_personal_presupuestado_en_la_intendencia <dbl>,
#   part_act_econ <dbl>
```

Es fácil de observar que existen variables numéricas que contienen observaciones de tipo `character` representando a valores faltantes y, al mismo tiempo, también observamos la existencia de valores `NA`.

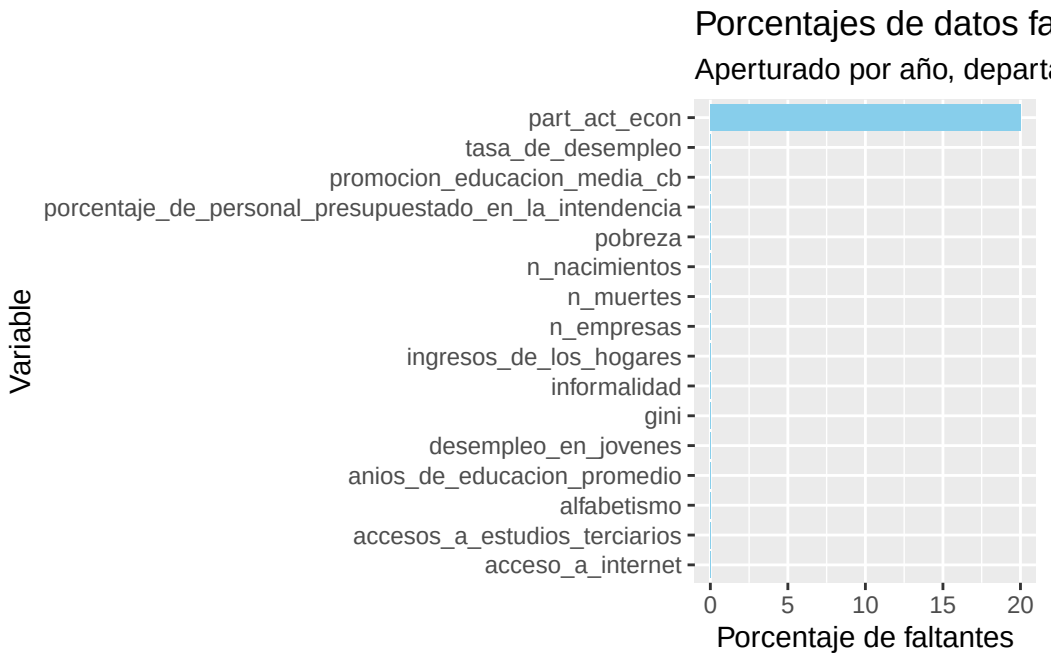
Entonces, nos preguntamos:

- ¿Cuál es el porcentaje de datos faltantes existe por cadá variable?

> **Note**
>
> Como se vio en las observaciones seleccionadas, existen datos faltantes representados con texto por lo que necesitaremos pre-procesar esas columnas y normalizar la representación de un dato faltante como `NA`. Luego formalizaremos este pre-procesamiento como parte del flujo pre-modelado.



- ¿Y si subimos un nivel de agregación ignorando el sector?



Podemos concluir que al estar tan desagregado nuestro Dataset los datos faltantes pasan a ser un problema a tener en cuenta. A partir de este punto realizaremos nuestro análisis con el

dataset aperturado por año, departamento y tamaño.

- ¿Estos NA pertenecen a algún período de tiempo particular?

```
# A tibble: 4 x 1
  anio
  <chr>
1 2019
2 2020
3 2021
4 2022
```

La serie de la participación en la actividad económica está incompleta para el período 2019-2022.

- ¿Cómo se comportan las distintas variables numéricas por tamaño de la empresa? ¿Cómo es su distribución? ¿Y el apartamiento?

```
$grandes
# A tibble: 16 x 10
   variable       min   p25 mediana  media    p75     max std_dev std_err     cv
   <chr>        <dbl> <dbl>   <dbl>  <dbl>  <dbl>   <dbl>   <dbl>   <dbl>  <dbl>
 1 n_empresas   1     5       7     45.7   15     1.08e+3 1.47e+2 8.73    322.
 2 n_nacimient~ 0     0       0      0.604  0     5.1 e+1 3.31e+0 0.196   548.
 3 n_muertes    0     0       0      0.653  0     3.3 e+1 2.86e+0 0.170   439.
 4 alfabetismo  0.725 0.866   0.9    0.895  0.928 9.78e-1 4.61e-2 0.00273   5.15
 5 accesos_a_e~ 0.058 0.15    0.182  0.215  0.236 7.62e-1 1.22e-1 0.00722  56.6
 6 anios_de_ed~ 0.141 0.304   0.349  0.357  0.397 6.78e-1 8.88e-2 0.00526  24.9
 7 promocion_e~ 0.168 0.471   0.564  0.568  0.672 9.15e-1 1.47e-1 0.00870  25.9
 8 pobreza      0.283 0.753   0.841  0.812  0.906 9.97e-1 1.27e-1 0.00753  15.6
 9 informalidad 0.148 0.42    0.511  0.498  0.584 7.89e-1 1.26e-1 0.00746  25.3
10 desempleo_e~ 0.214 0.511   0.579  0.577  0.654 8.86e-1 1.12e-1 0.00665  19.5
11 gini         0.288 0.522   0.592  0.589  0.666 8.36e-1 1.10e-1 0.00649  18.6
12 acceso_a_in~ 0.043 0.334   0.477  0.458  0.595 9.48e-1 2.05e-1 0.0122   44.9
13 ingresos_de~ 0.083 0.238   0.315  0.329  0.384 7.9 e-1 1.40e-1 0.00831  42.7
14 tasa_de_des~ 0.149 0.605   0.68   0.675  0.76  9.75e-1 1.28e-1 0.00761  19.0
15 porcentaje_~ 0.009 0.264   0.379  0.387  0.484 8.32e-1 1.73e-1 0.0102   44.6
16 part_act_ec~ 0.7   1.6     2.1    5.26   2.9   5.11e+1 1.08e+1 0.642   206.

$medianas
# A tibble: 16 x 10
   variable    min    p25 mediana   media     p75     max std_dev std_err     cv
```

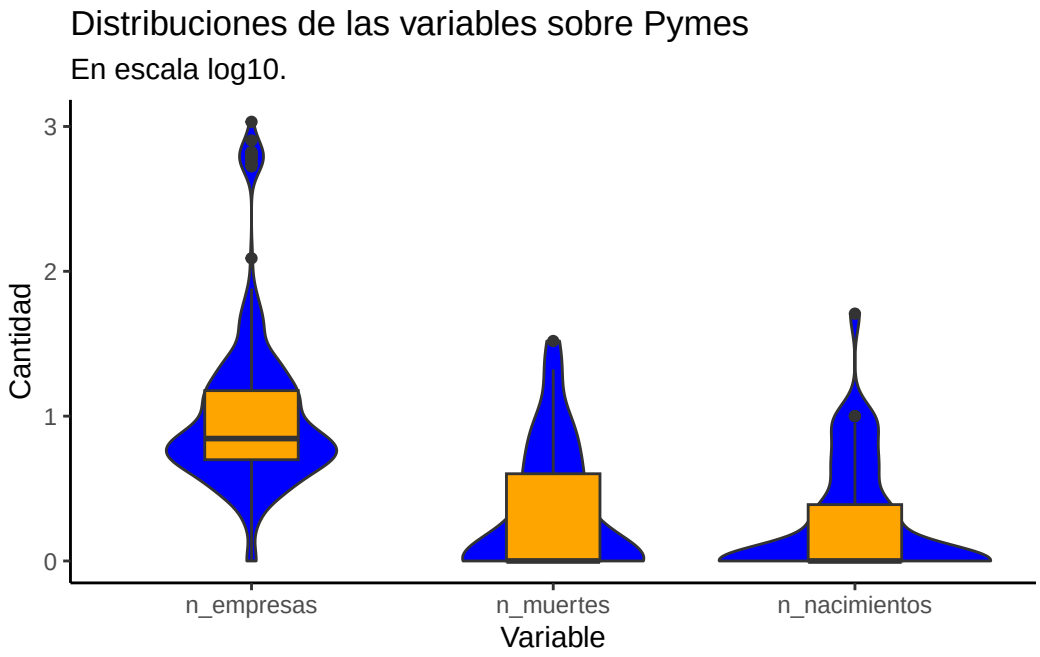
```
   <chr>     <dbl>  <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>  <dbl>
 1 n_empre~ 11     47      67     278.    130     5.33e+3 7.38e+2 4.37e+1 266.
 2 n_nacim~  0      0       1       5.07    3     1.34e+2 1.48e+1 8.74e-1 291.
 3 n_muert~  0      0       1       4.56    3     8.6 e+1 1.26e+1 7.45e-1 275.
 4 alfabet~  0.725  0.866   0.9     0.895   0.928 9.78e-1 4.61e-2 2.73e-3   5.15
 5 accesos~  0.058  0.15    0.182   0.215   0.236 7.62e-1 1.22e-1 7.22e-3  56.6
 6 anios_d~  0.141  0.304   0.349   0.357   0.397 6.78e-1 8.88e-2 5.26e-3  24.9
 7 promoci~  0.168  0.471   0.564   0.568   0.672 9.15e-1 1.47e-1 8.70e-3  25.9
 8 pobreza   0.283  0.753   0.841   0.812   0.906 9.97e-1 1.27e-1 7.53e-3  15.6
 9 informa~  0.148  0.42    0.511   0.498   0.584 7.89e-1 1.26e-1 7.46e-3  25.3
10 desempl~  0.214  0.511   0.579   0.577   0.654 8.86e-1 1.12e-1 6.65e-3  19.5
11 gini      0.288  0.522   0.592   0.589   0.666 8.36e-1 1.10e-1 6.49e-3  18.6
12 acceso_~  0.043  0.334   0.477   0.458   0.595 9.48e-1 2.05e-1 1.22e-2  44.9
13 ingreso~  0.083  0.238   0.315   0.329   0.384 7.9 e-1 1.40e-1 8.31e-3  42.7
14 tasa_de~  0.149  0.605   0.68    0.675   0.76  9.75e-1 1.28e-1 7.61e-3  19.0
15 porcent~  0.009  0.264   0.379   0.387   0.484 8.32e-1 1.73e-1 1.02e-2  44.6
16 part_ac~  0.7    1.6     2.1     5.26    2.9   5.11e+1 1.08e+1 6.42e-1 206.

$micro
# A tibble: 16 x 10
   variable          min     p25 mediana   media     p75     max std_dev std_err
   <chr>           <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
 1 n_empresas    893     2.03e+3 2.84e+3 7.45e+3 4.36e+3 8.43e+4 1.52e+4 9.02e+2
 2 n_nacimientos   0     3.73e+2 5.61e+2 1.37e+3 8.54e+2 1.78e+4 2.80e+3 1.66e+2
 3 n_muertes      94     3.18e+2 4.75e+2 1.20e+3 7.33e+2 1.41e+4 2.46e+3 1.45e+2
 4 alfabetismo     0.725 8.66e-1 9   e-1 8.95e-1 9.28e-1 9.78e-1 4.61e-2 2.73e-3
 5 accesos_a_es~   0.058 1.5 e-1 1.82e-1 2.15e-1 2.36e-1 7.62e-1 1.22e-1 7.22e-3
 6 anios_de_edu~   0.141 3.04e-1 3.49e-1 3.57e-1 3.97e-1 6.78e-1 8.88e-2 5.26e-3
 7 promocion_ed~   0.168 4.71e-1 5.64e-1 5.68e-1 6.72e-1 9.15e-1 1.47e-1 8.70e-3
 8 pobreza         0.283 7.53e-1 8.41e-1 8.12e-1 9.06e-1 9.97e-1 1.27e-1 7.53e-3
 9 informalidad    0.148 4.2 e-1 5.11e-1 4.98e-1 5.84e-1 7.89e-1 1.26e-1 7.46e-3
10 desempleo_en~   0.214 5.11e-1 5.79e-1 5.77e-1 6.54e-1 8.86e-1 1.12e-1 6.65e-3
11 gini            0.288 5.22e-1 5.92e-1 5.89e-1 6.66e-1 8.36e-1 1.10e-1 6.49e-3
12 acceso_a_int~   0.043 3.34e-1 4.77e-1 4.58e-1 5.95e-1 9.48e-1 2.05e-1 1.22e-2
13 ingresos_de_~   0.083 2.38e-1 3.15e-1 3.29e-1 3.84e-1 7.9 e-1 1.40e-1 8.31e-3
14 tasa_de_dese~   0.149 6.05e-1 6.8 e-1 6.75e-1 7.6 e-1 9.75e-1 1.28e-1 7.61e-3
15 porcentaje_d~   0.009 2.64e-1 3.79e-1 3.87e-1 4.84e-1 8.32e-1 1.73e-1 1.02e-2
16 part_act_econ   0.7   1.6 e+0 2.1 e+0 5.26e+0 2.9 e+0 5.11e+1 1.08e+1 6.42e-1
# i 1 more variable: cv <dbl>

$pequeñas
# A tibble: 16 x 10
   variable          min     p25 mediana   media     p75     max std_dev std_err
```

```
   <chr>           <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
 1 n_empresas    115     301     408     1.23e+3 621     1.50e+4 2.75e+3 1.63e+2
 2 n_nacimientos   0      10      16     4.99e+1  31     6.87e+2 1.14e+2 6.74e+0
 3 n_muertes       0      10      17     4.93e+1  31     6.57e+2 1.04e+2 6.18e+0
 4 alfabetismo     0.725   0.866   0.9   8.95e-1   0.928 9.78e-1 4.61e-2 2.73e-3
 5 accesos_a_es~   0.058   0.15    0.182 2.15e-1   0.236 7.62e-1 1.22e-1 7.22e-3
 6 anios_de_edu~   0.141   0.304   0.349 3.57e-1   0.397 6.78e-1 8.88e-2 5.26e-3
 7 promocion_ed~   0.168   0.471   0.564 5.68e-1   0.672 9.15e-1 1.47e-1 8.70e-3
 8 pobreza         0.283   0.753   0.841 8.12e-1   0.906 9.97e-1 1.27e-1 7.53e-3
 9 informalidad    0.148   0.42    0.511 4.98e-1   0.584 7.89e-1 1.26e-1 7.46e-3
10 desempleo_en~   0.214   0.511   0.579 5.77e-1   0.654 8.86e-1 1.12e-1 6.65e-3
11 gini            0.288   0.522   0.592 5.89e-1   0.666 8.36e-1 1.10e-1 6.49e-3
12 acceso_a_int~   0.043   0.334   0.477 4.58e-1   0.595 9.48e-1 2.05e-1 1.22e-2
13 ingresos_de_~   0.083   0.238   0.315 3.29e-1   0.384 7.9 e-1 1.40e-1 8.31e-3
14 tasa_de_dese~   0.149   0.605   0.68  6.75e-1   0.76  9.75e-1 1.28e-1 7.61e-3
15 porcentaje_d~   0.009   0.264   0.379 3.87e-1   0.484 8.32e-1 1.73e-1 1.02e-2
16 part_act_econ   0.7     1.6     2.1   5.26e+0   2.9   5.11e+1 1.08e+1 6.42e-1
# i 1 more variable: cv <dbl>
```

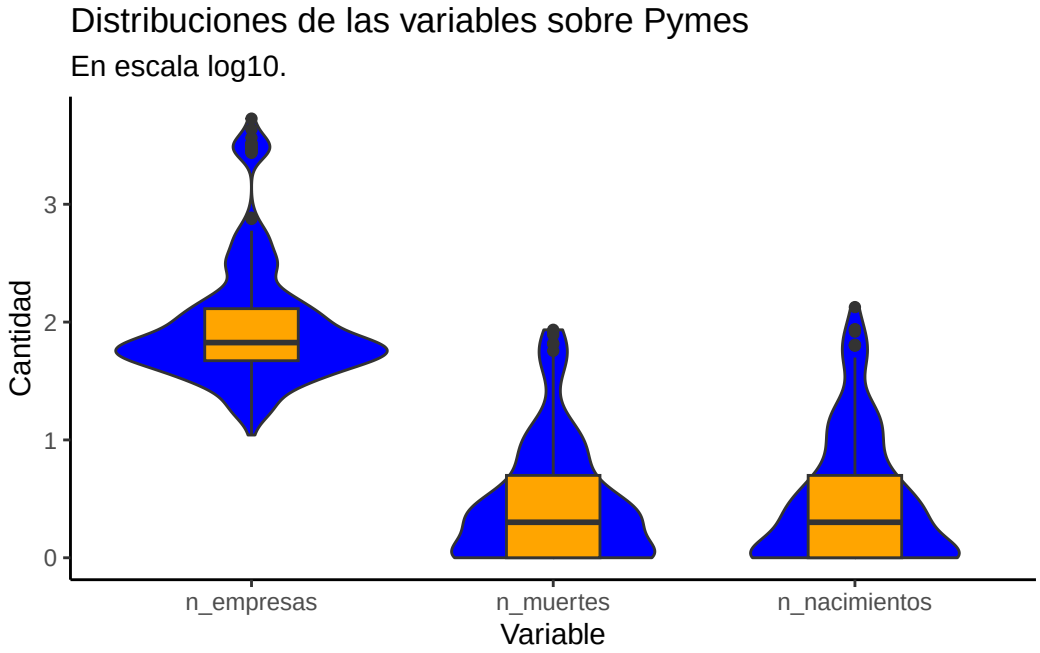
Visualmente:

- Variables sobre Pymes

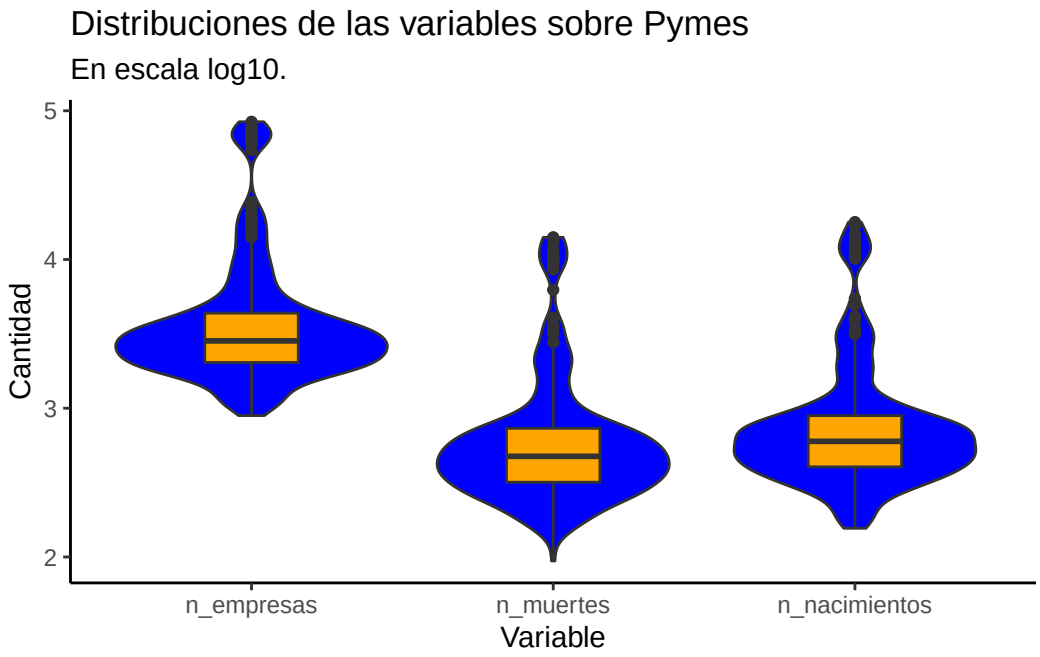
```
$grandes
```



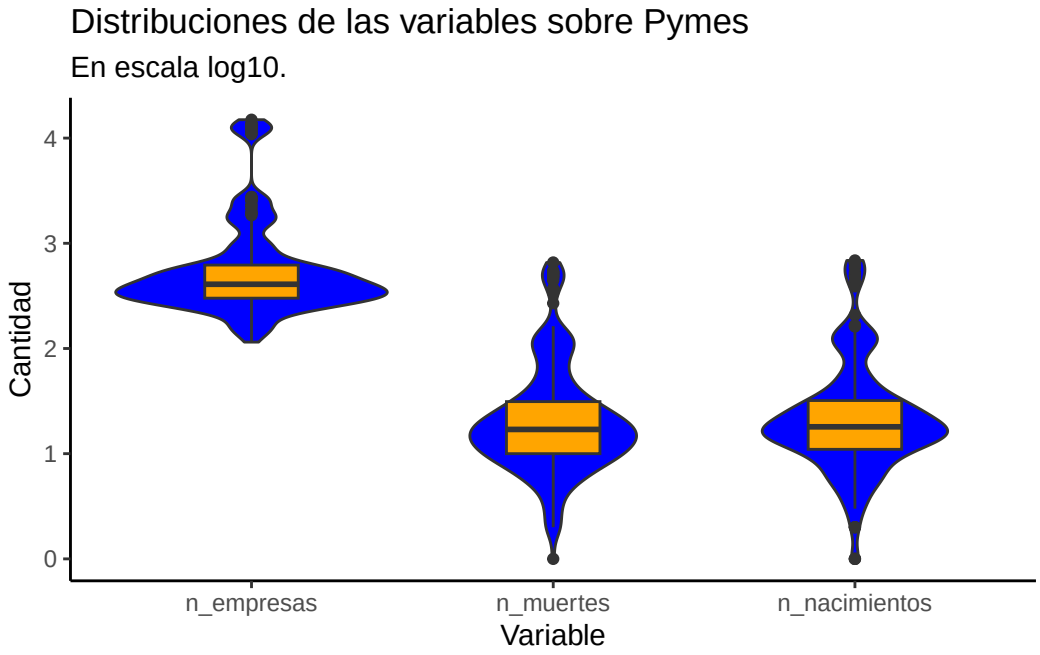

```
$medianas
```




```
$micro
```

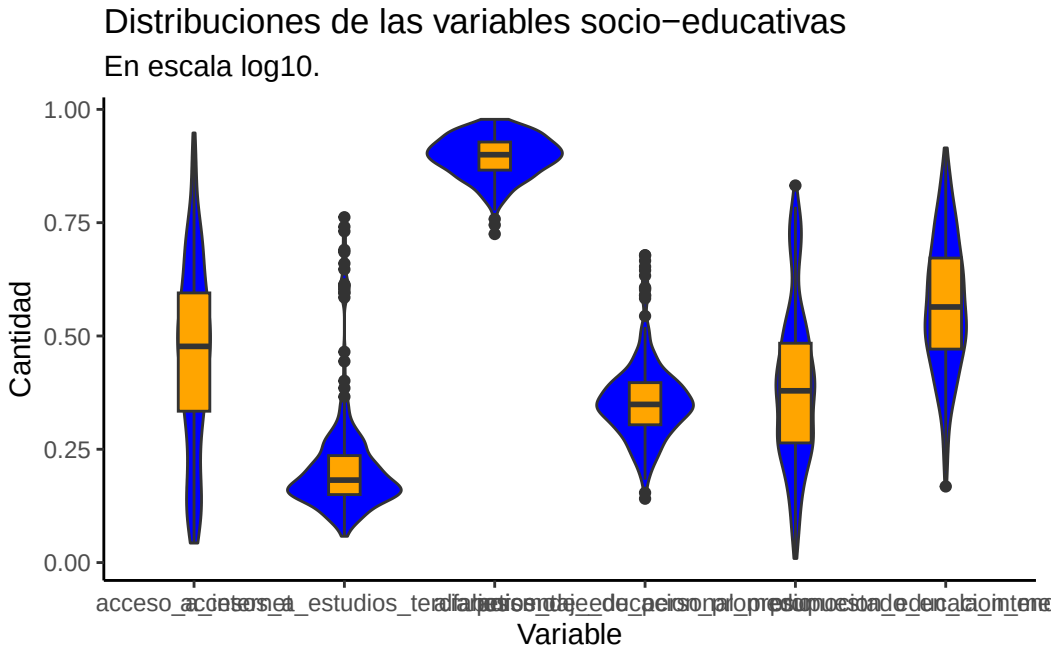



```
$pequeñas
```

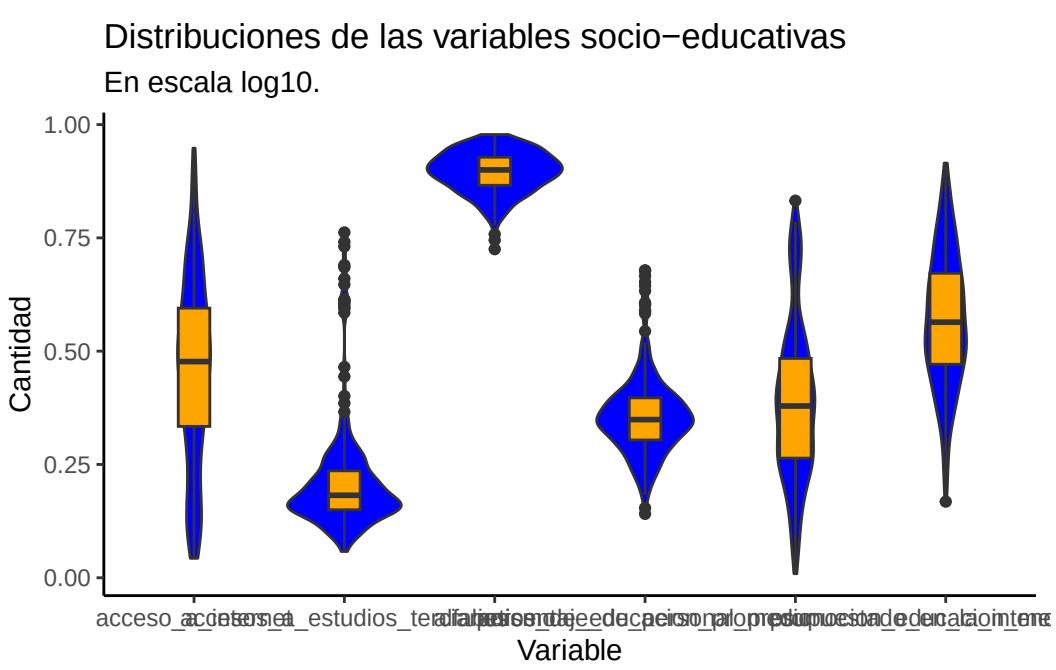


- Variables sobre socio-educativas

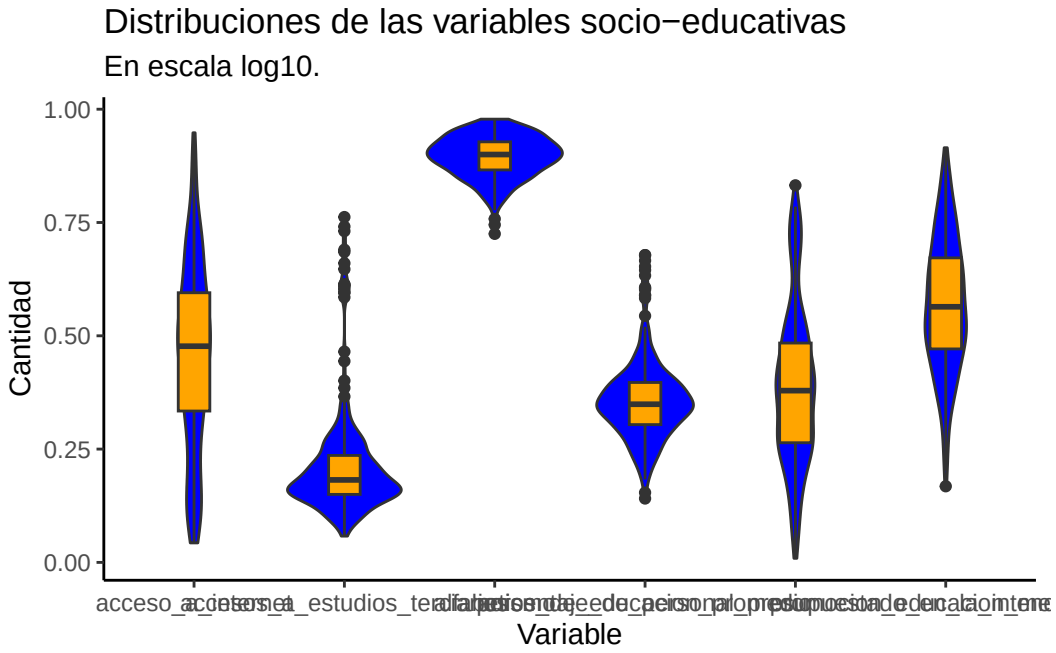
```
$grandes
```



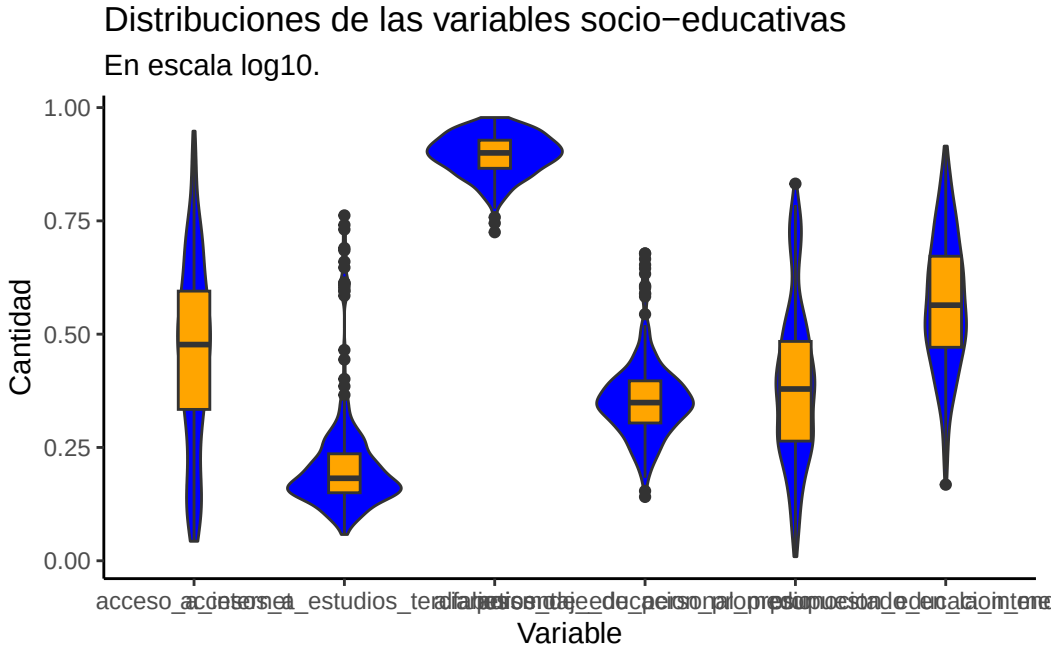

```
$medianas
```



```
$micro
```

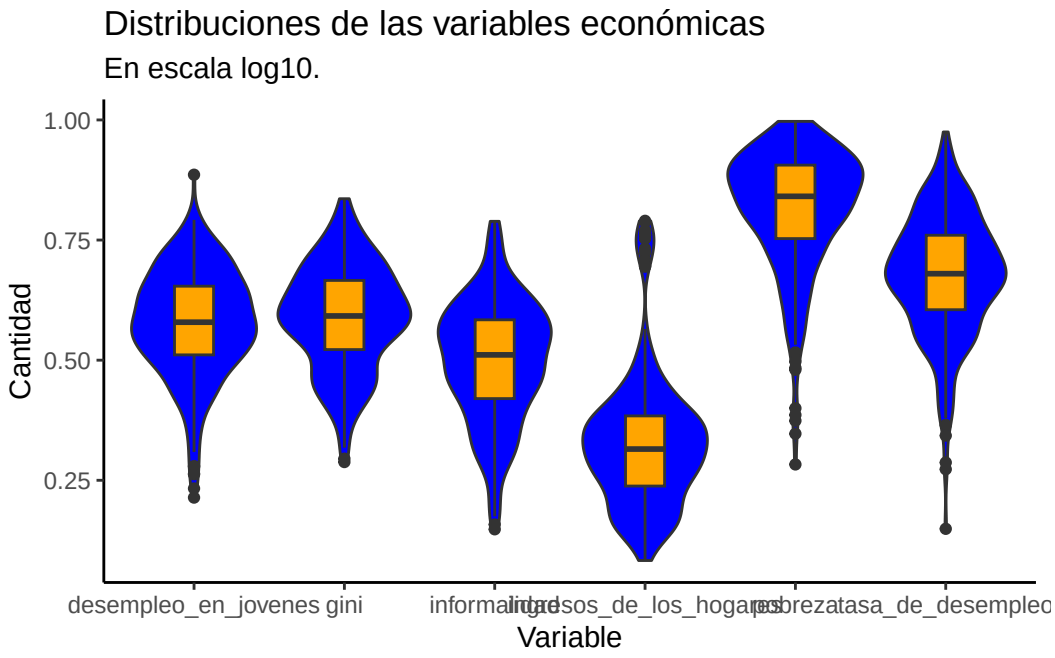



```
$pequeñas
```

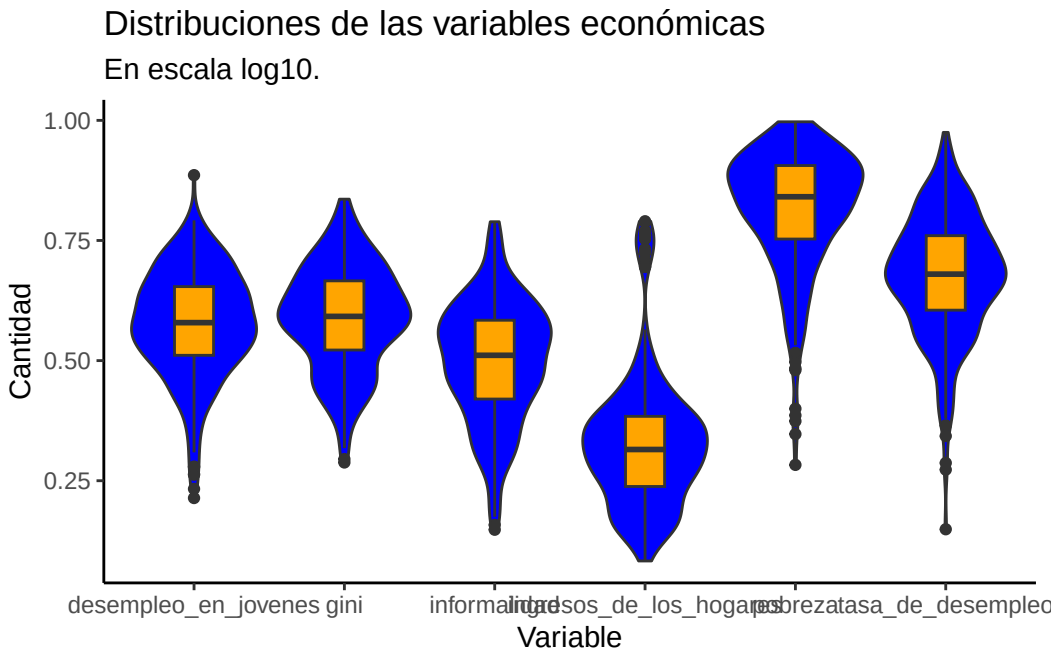


- Variables económicas:

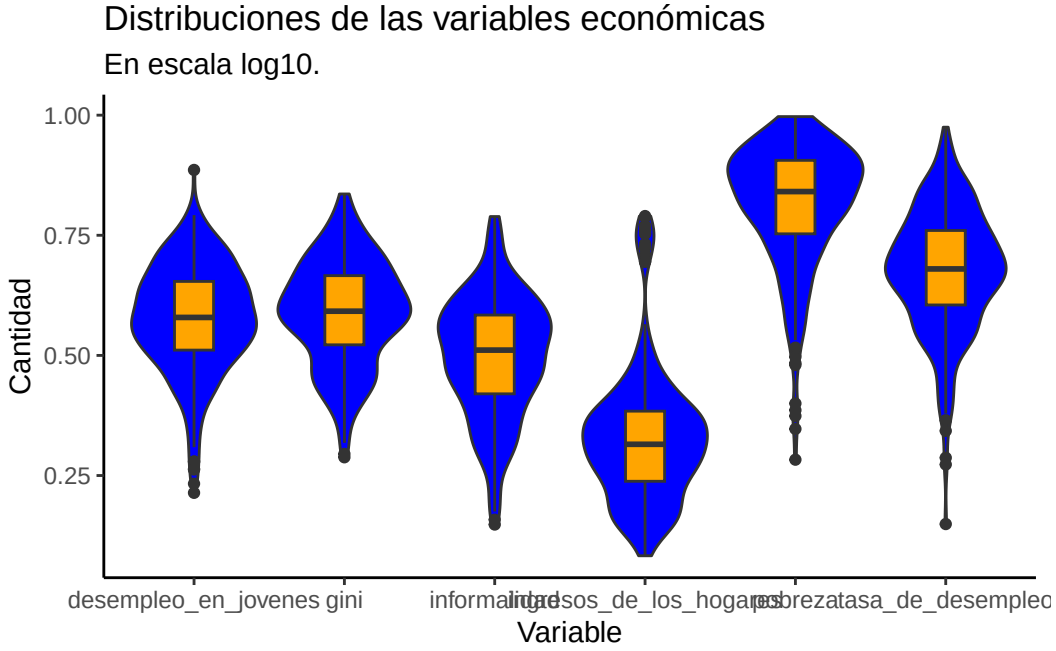
```
$grandes
```



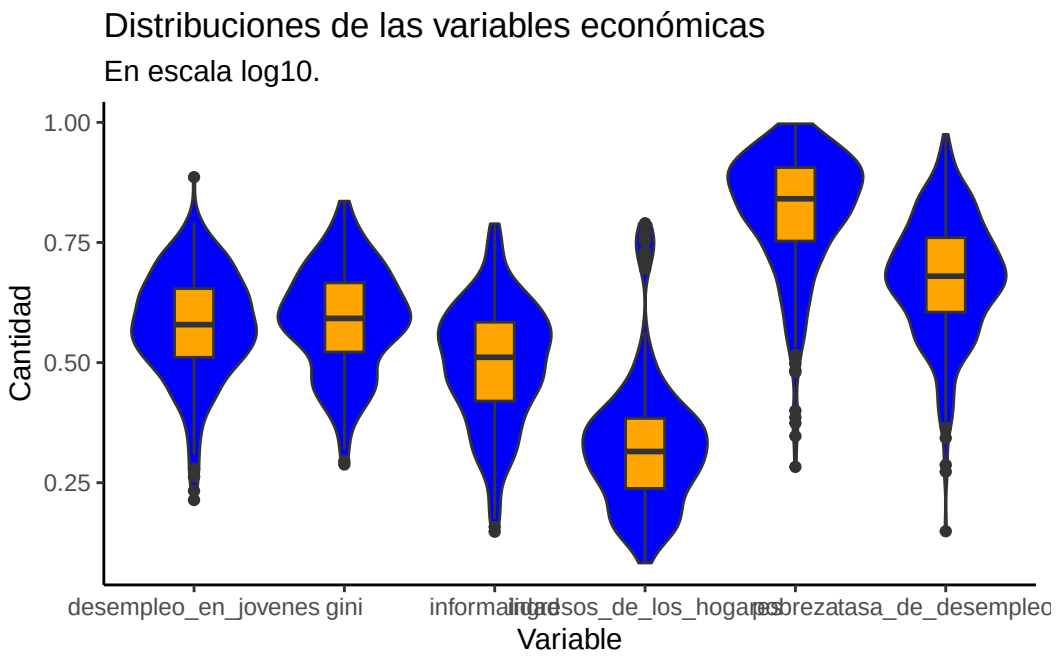

```
$medianas
```




```
$micro
```

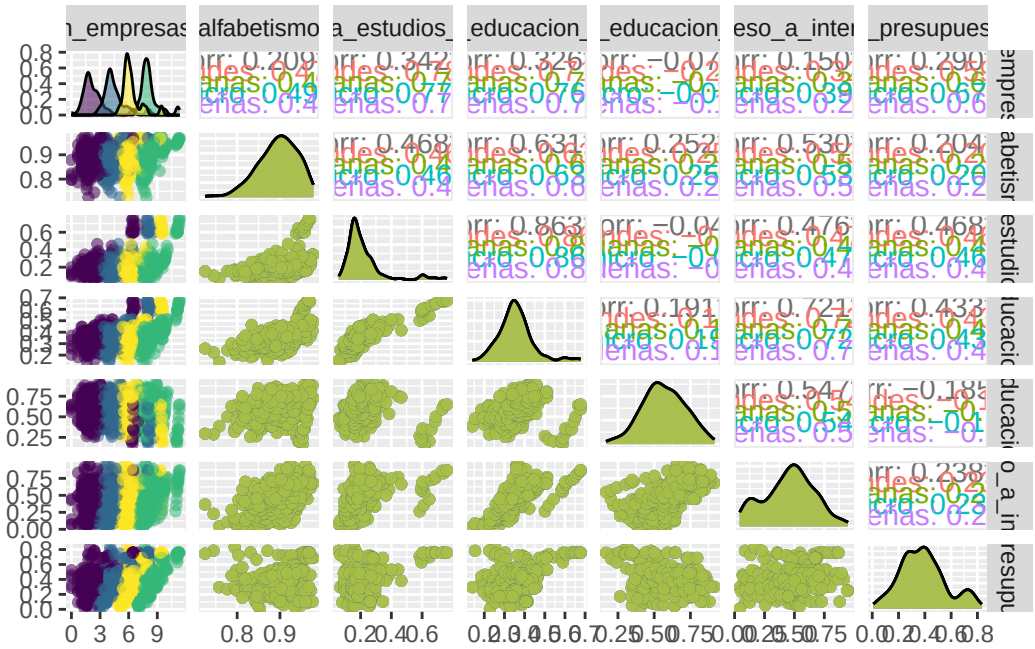



```
$pequeñas
```

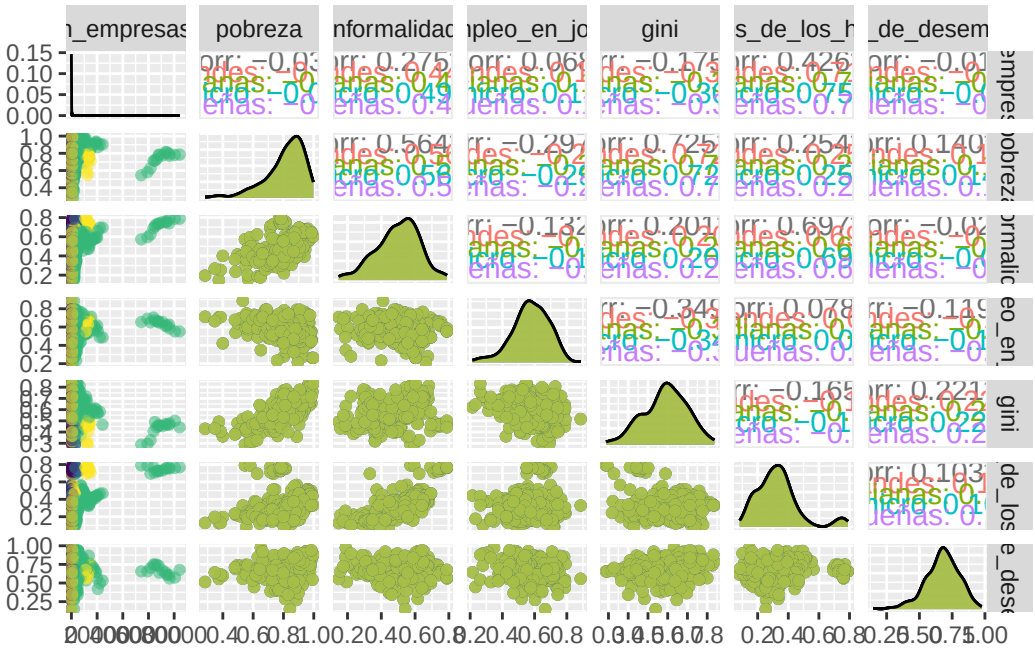


Se puede observar que que la mayoría de variables numéricas tienden a tener bastantes datos atípicos atípicos lo que genera que las medidas de apartamiento se vean afectadas y haya una diferencia importante entre las medianas y medias de cada variable. A su vez, a excepción de las variables sobre pymes, tienden a tener una forma parecida a una normal con cola alargada (positiva o negativa depende mucho del caso).

- ¿Cómo se relacionan las distintas variables numéricas con la cantidad de empresas?
  - Con las variables socio-educativas



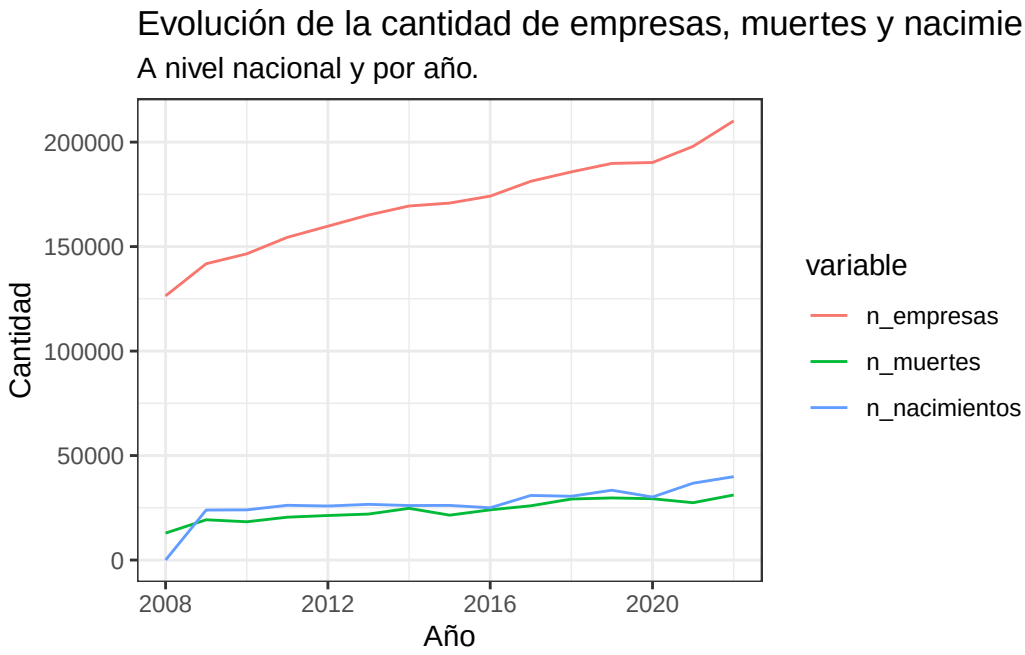
- Con las variables economicas:



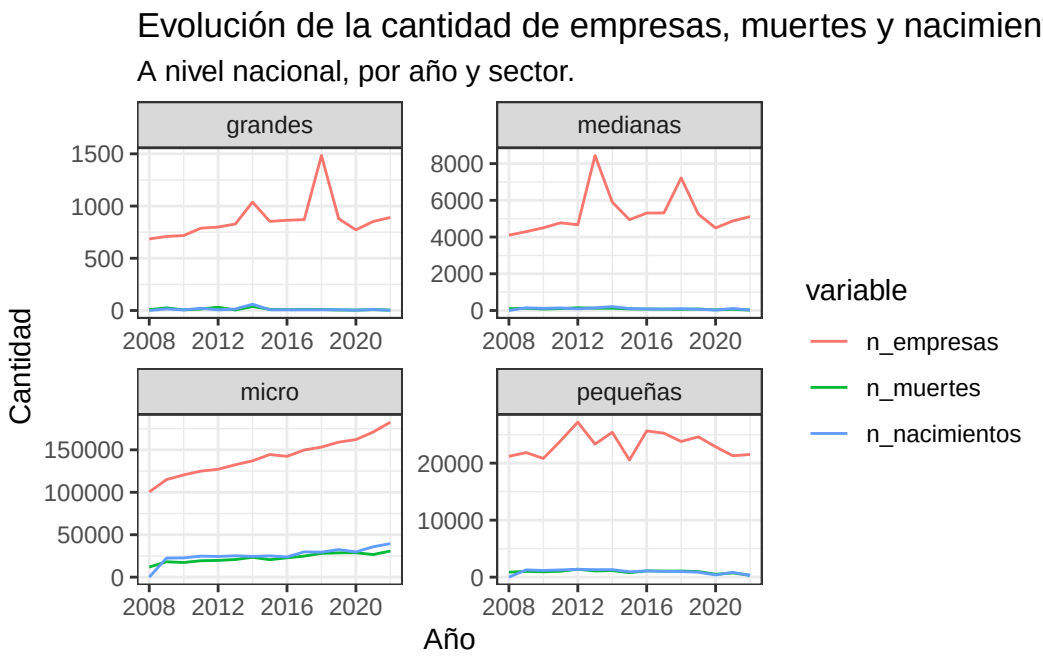
Se puede observar claramente que las variables socio-educativas y económicas son a nivel departamento y que el tamaño de empresa cambia la dispersión de la cantidad de empresas

contra las otras variables.

- ¿Cómo evolucionan los variables de pymes a lo largo de los años?



- ¿Y por tamaño de la empresa?



Se tiene que, en general, las muertes y nacimientos se mantienen con poca variación (habría que comprobar que no haya una variable oculta) y que éxisten outliers para las empresas de mayor tamaño mientras que las micro siguen una tendencia creciente pero no así las pequeñas que son más variables en el tiempo.

Preprocesamiento de datos

Como se pudo observar durante la fase del EDA, nuestro dataset crudo tiene problema de consistencia e incompletitud en los datos que, en ésta sección, abordaremos para transformarlo hacia un dataset que nos permita realizar una predicción lo más confiable posible.

Débido a esto es que realizaremos las siguientes transformaciones antes de modelar:

- Utilizaremos año, departamento y tamaño como apertura.
- Transformaremos todos los datos faltantes a NA.
- Redondearemos a 3 decimales las variables numéricas.
- Utilizaremos la serie sin datos faltantes (2008-2018)

Modelado

Es de nuestro interés la creación de un modelo predictivo de la variable `n_empresas` y, para cumplir con dicho objetivo, utilizaremos el resto de variables como predictores de `n_empresas`

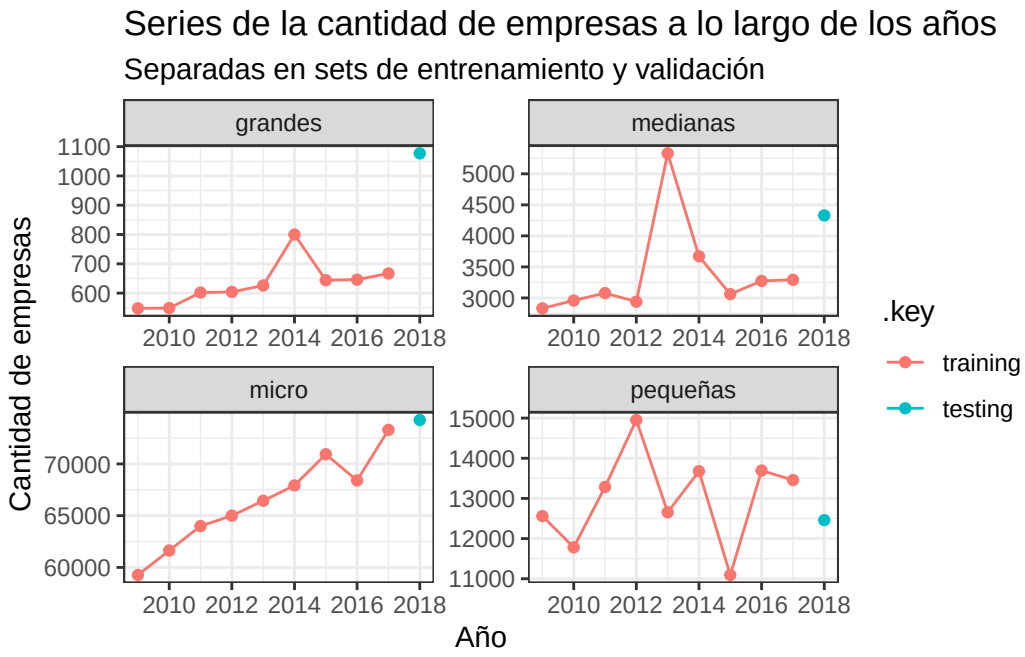
tratando de modelar cómo la situación educacional, social y económica de cada departamento influye en la creación de Pymes.

Empezaremos estableciendo un punto de partida con un modelo "baseline" que queremos superar con modelos más complejos que sepan inferir las relaciones complejas en éste problema multidimensional.

Separación en datos de entrenamiento y de validación

En un problema de predicción, la separación de entrenamiento difiere un poco de lo que se hace para una regresión ya que al querer generar un modelo que predice el futuro nuestra validación reside en que tan bien éste predice en comparación a un evento que realmente ocurrió.

Una práctica común al momento de hacer ésta separación de datos es la de separar basados en tu horizonte de predicción (cuantos puntos a futuro se quiere predecir). Utilizando ésta referencia es que utilizaremos el 2017 como punto de separación



Regresión lineal

Para nuestro modelo base, eligiremos una regresión lineal como nuestro modelo base. Éste es un modelo básico al momento de capturar relaciones complejas o no lineales ya que se basa en

la estimación de nuestra variable de interés (y) dada la sumatoria del resto de predictores (x_i) por un coeficiente (b_i) y un error (

$$\varepsilon_i$$

), tal que:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Árboles

Entrenaremos dos tipos distintos de árboles para intentar capturar eficientemente las variables de mayor peso al momento de estimar la cantidad de empresas.

Para ello, primero haremos una búsqueda de hiperparámetros y luego seleccionaremos el mejor modelo en base al `RMSE` como métrica de minimización elegida. Ésta métrica es elegida en base a su poder de interpretabilidad al estar en las mismas unidades que la variable a predecir.

Bosque aleatorio

Random Forest es una técnica de ensamble basada en árboles de decisión y en el principio de bagging. Este método entrena múltiples árboles de decisión en subconjuntos aleatorios del conjunto de datos, y combina sus predicciones mediante promedio, lo cual reduce significativamente la varianza del modelo individual.

Random Forest es particularmente adecuado en este contexto porque:

- Es robusto frente al sobreajuste, especialmente cuando hay una alta cardinalidad en variables categóricas como el departamento.
- Captura bien interacciones y no linealidades.
- Su estructura permite interpretar la importancia relativa de las variables predictoras, lo cual nos ayuda a entender la estimación.

XGBoost

XGBoost (Extreme Gradient Boosting), una técnica de aprendizaje supervisado basada en árboles de decisión y en el principio de boosting. Este algoritmo construye múltiples árboles de forma secuencial, donde cada árbol intenta corregir los errores del anterior. A diferencia de la regresión lineal, XGBoost permite capturar relaciones no lineales y complejas interacciones entre las variables predictoras.

Es un modelo que:

- Robusto al overfitting

- Capta relaciones no lineales

Entrenamiento

```
# Modeltime Table
# A tibble: 6 x 5
  .model_id .model     .model_desc          .type .calibration_data
      <int> <list>     <chr>                <chr> <list>
1         1 <workflow> RECIPE_RAND_FOREST_1 Test  <tibble [76 x 4]>
2         2 <workflow> RECIPE_RAND_FOREST_2 Test  <tibble [76 x 4]>
3         3 <workflow> RECIPE_RAND_FOREST_3 Test  <tibble [76 x 4]>
4         4 <workflow> RECIPE_RAND_FOREST_4 Test  <tibble [76 x 4]>
5         5 <workflow> RECIPE_RAND_FOREST_5 Test  <tibble [76 x 4]>
6         6 <workflow> RECIPE_RAND_FOREST_6 Test  <tibble [76 x 4]>
```

Evaluación

A partir de los modelos entrenados calcularemos las métricas de error poniendo foco en el RMSE|

```
# A tibble: 54 x 9
   .model_id .model_desc           .type   mae   mape   mase smape  rmse   rsq
       <int> <chr>                 <chr> <dbl>  <dbl>  <dbl> <dbl> <dbl> <dbl>
 1         1 RECIPE_RAND_FOREST_1  Test  2089. 1140.  0.517  123.  7207. 0.367
 2         2 RECIPE_RAND_FOREST_2  Test   458.   50.4 0.113   43.9 1097. 0.992
 3         3 RECIPE_RAND_FOREST_3  Test   376.   75.0 0.0930  35.5 1163. 0.983
 4         4 RECIPE_RAND_FOREST_4  Test  1857.   91.2 0.459   65.9 7400. 0.354
 5         5 RECIPE_RAND_FOREST_5  Test   412.  118.  0.102   47.6 1261. 0.996
 6         6 RECIPE_RAND_FOREST_6  Test   463.   37.9 0.115   38.5 1243. 0.987
 7         7 RECIPE_RAND_FOREST_7  Test  1080.  109.  0.267   71.8 2355. 0.940
 8         8 RECIPE_RAND_FOREST_8  Test  2019.  217.  0.499   62.5 7010. 0.372
 9         9 RECIPE_RAND_FOREST_9  Test  1364.  579.  0.337   48.6 3371. 0.860
10        10 RECIPE_RAND_FOREST_10 Test  1491.  968.  0.369  101.  4538. 0.846
# i 44 more rows
```

Selección del mejor modelo

De XGBoost vimos que el mejor modelo es:

```
# A tibble: 1 x 3
  learn_rate tree_depth trees
       <dbl>      <int> <int>
1        0.1          8  1000
```

Y para el random forest:

```
# A tibble: 1 x 3
   mtry min_n trees
  <int> <int> <int>
1    10     2  2000
```

Entonces reentrenemos un único modelo, calculando predicciones e importancia de cada feature al momento de decidir.

Predicciones

```
[[1]]
# A tibble: 76 x 4
   anio       .actual .prediction .residuals
   <date>       <dbl>       <dbl>      <dbl>
 1 2018-01-01       3     -264.        267.
 2 2018-01-01      68       -4.80       72.8
 3 2018-01-01    2270     3254.       -984.
 4 2018-01-01     303      575.       -272.
 5 2018-01-01     123       11.1       112.
 6 2018-01-01     751      327.        424.
 7 2018-01-01   19238    20384.      -1146.
 8 2018-01-01    2625     1320.       1305.
 9 2018-01-01       9     -266.        275.
10 2018-01-01      87      -17.8       105.
# i 66 more rows

[[2]]
# A tibble: 76 x 4
   anio       .actual .prediction .residuals
   <date>       <dbl>       <dbl>      <dbl>
 1 2018-01-01       3        8.45      -5.45
 2 2018-01-01      68       57.7        10.3
 3 2018-01-01    2270     2770.       -500.
 4 2018-01-01     303      330.        -26.9
```

```
 5 2018-01-01     123      686.      -563.
 6 2018-01-01     751     1066.      -315.
 7 2018-01-01   19238    16720.      2518.
 8 2018-01-01    2625     3059.      -434.
 9 2018-01-01       9       13.4       -4.41
10 2018-01-01      87       72.6       14.4
# i 66 more rows

[[3]]
# A tibble: 76 x 4
   anio       .actual .prediction .residuals
   <date>       <dbl>       <dbl>      <dbl>
 1 2018-01-01       3        3.72    -0.720
 2 2018-01-01      68       44.7      23.3
 3 2018-01-01    2270     2545.     -275.
 4 2018-01-01     303      299.       4.08
 5 2018-01-01     123       51.5     71.5
 6 2018-01-01     751      575.      176.
 7 2018-01-01   19238    18600.      638.
 8 2018-01-01    2625     2575.      50.2
 9 2018-01-01       9        8.93     0.0738
10 2018-01-01      87       80.0      7.03
# i 66 more rows
```

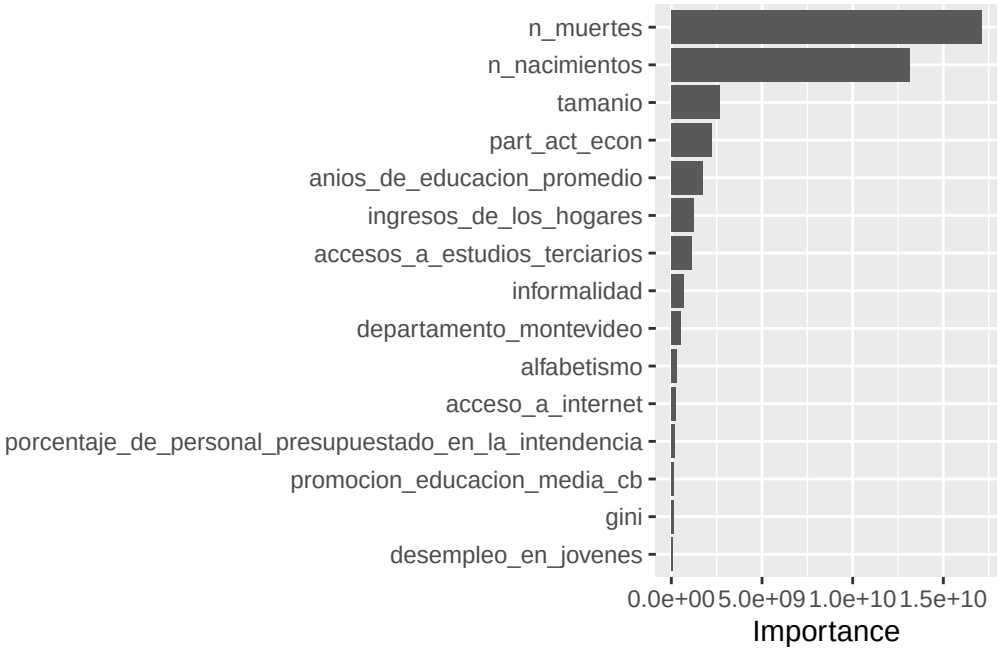
Re-evaluación

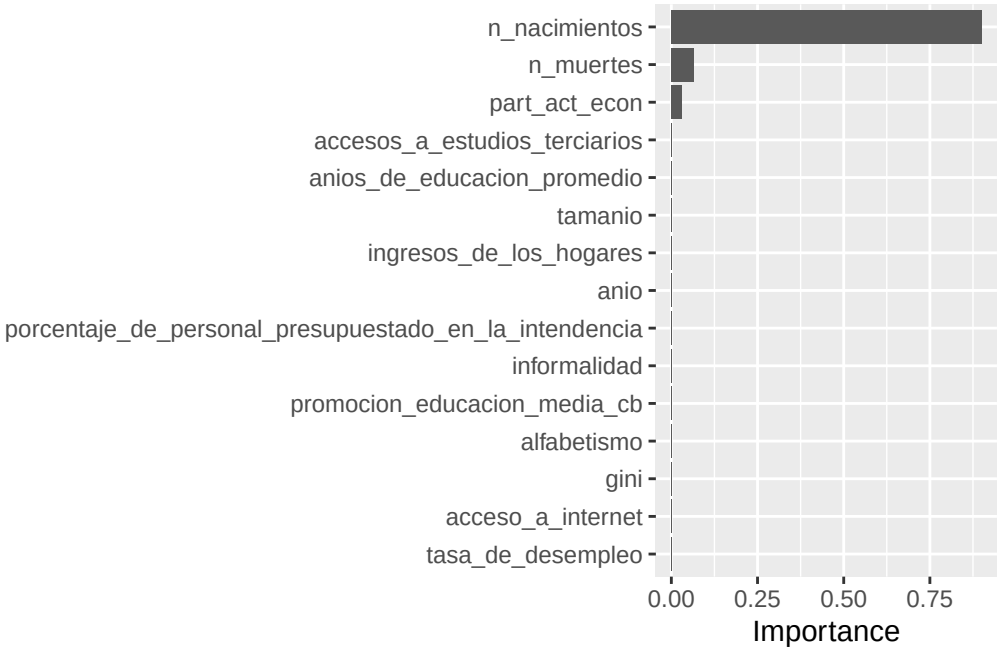
Reevaluamos junto a las métricas de la regresión lineal.

```
# A tibble: 3 x 9
  .model_id .model_desc .type   mae  mape   mase smape  rmse   rsq
      <int> <chr>       <chr> <dbl> <dbl>  <dbl> <dbl> <dbl> <dbl>
1         1 GLMNET      Test   394. 514.  0.0975  87.7  789. 0.992
2         2 RANGER      Test   297.  35.4 0.0736  25.6 1237. 0.998
3         3 XGBOOST     Test   163.  21.2 0.0402  25.1  420. 0.999
```

Importancia de las variables en los árboles

Es interesante el poder explicar como se llegá a una conclusión (una predicción en este caso), para eso visualizaremos que importancia, a través del "impurity" (en el caso del random forest), se le dio a cada variable.


n_muertes
n_nacimientos
tamanio
part_act_econ
anios_de_educacion_promedio
ingresos_de_los_hogares
accesos_a_estudios_terciarios
informalidad
departamento_montevideo
alfabetismo
acceso_a_internet
porcentaje_de_personal_presupuestado_en_la_intendencia
promocion_educacion_media_cb
gini
desempleo_en_jovenes
0.0e+00
5.0e+09
1.0e+10
1.5e+10
Importance


n_nacimientos
n_muertes
part_act_econ
accesos_a_estudios_terciarios
anios_de_educacion_promedio
tamanio
ingresos_de_los_hogares
anio
porcentaje_de_personal_presupuestado_en_la_intendencia
informalidad
promocion_educacion_media_cb
alfabetismo
gini
acceso_a_internet
tasa_de_desempleo
0.00
0.25
0.50
0.75
Importance

Referencias

Agencia Nacional de Desarrollo (ANDE). 2024. "Indicadores Demográficos – Monitor MIPYMES." https://monitor.ande.org.uy/indicadoresemp.aspx.

Observatorio Territorio Uruguay – Oficina de Planeamiento y Presupuesto (OPP). 2018. "Actividad Económica Departamental: Participación Porcentual En La Actividad Económica Nacional (2008–2018)." Conjunto de datos. Oficina de Planeamiento y Presupuesto (Uruguay). https://www.opp.gub.uy/.

Rodríguez Miranda, Andrés, Claudio Vial Cossani, Ignacio Centurión, and Mariana Pérez. 2024. "Índice de Desarrollo Regional Uruguay 2006–2022. IDERE-UY. Informe 2024." Informe técnico. Montevideo: Facultad de Ciencias Económicas y de Administración, Universidad de la República. https://desarrolloterritorial.idere.ei.udelar.edu.uy/.