DEPARTAMENTO DE COMPUTACION

Facultad de Ciencias Exactas y Naturales - UBA

Trabajo Práctico 1

Los salarios en el sector de producción orgánica del país

October 29, 2023 Laboratorio de Datos

Integrante	LU	Correo electrónico
Guibaudo, Camila	682/17	cami_sol_guibaudo@hotmail.com
Dembling, Ariel	408/92	arieldembling@gmail.com
Suarez, Antony	792/21	sebastsuar@gmail.com

Facultad de Ciencias Exactas y Naturales
Universidad de Buenos Aires
Ciudad Universitaria - (Pabellón I/Planta Baja)
Intendente Güiraldes 2610 - C1428EGA
Ciudad Autónoma de Buenos Aires - Rep. Argentina
Tel/Fax: (++54 +11) 4576-3300
http://www.exactas.uba.ar

1 Resumen

2 Introducción

En el presente informe se documenta el desarrollo del Trabajo Práctico 1 de la materia 'Laboratorio de Datos'. El objetivo principal de este trabajo es determinar si existe cierta relación entre el desarrollo de la actividad de los Operadores Orgánicos Certificados y el salario promedio que perciben los trabajadores del sector privado en cada departamento de las provincias argentinas. Para ello se trabajó analizando los siguientes conjuntos de datos:

- **Padrón de Operadores Orgánicos Certificados**, cuyo responsable es la Dirección de Agroalimentos - Producción Orgánica, y fue obtenido del sitio que se detalla a continuación: https://datos.magyp.gob.ar/dataset/padron-de-operadores-organicos-certificados.
- **Salarios del sector privado**. Fuente de datos que contiene el salario bruto mediano de los trabajadores registrados del sector privado, por departamento/partido y clase, con frecuencia mensual y desde 2014. El responsable de dichos datos es el Ministerio de Desarrollo Productivo. Unidad Gabinete de Asesores. Dirección Nacional de Estudios para la Producción. Los datos fueron obtenidos de: https://www.datos.gob.ar/fa_IR/dataset/produccion-salarios-por-departamentopartido -sector-actividad/archivo/proc d008-42fa-a9d7-8a1bb26d04ab.
- **Listado de las localidades censales según la base de datos censales del INSTITUTO NACIONAL DE ESTADÍSTICA Y CENSOS (INDEC)**, cuyo responsable es la Jefatura de Gabinete de Ministros. Secretaría de Innovación Pública. Subsecretaría de Servicios y País Digital. Dicha fuente se puede obtener de: https://datos.gob.ar/ar/dataset/jgm-servicio-normalizacion-datos-geograficos/archivo/j gm_8.12.
- **Diccionario de departamentos**. El responsable de dichos datos es el Ministerio de Desarrollo Productivo - Unidad Gabinete de Asesores - Dirección Nacional de Estudios para la Producción. La fuente contiene los códigos utilizados por el INDEC para caracterizar los departamentos/partidos y provincias, con su correspondiente descripción. En el caso del código de CABA es un un código ficticio. Dicha fuente se puede obtener de: https://datos.produccion.gob.ar/dataset/puestos-de-trabajo-por-departamento-partido -y-sector-de-actividad/archivo/ 0205-417a-bf20-76d34dbe184b.
- **Diccionario de clases**, cuyo responsable es el Ministerio de Desarrollo Productivo - Unidad Gabinete de Asesores - Dirección Nacional de Estudios para la Producción. Dicha fuente contiene los nomencladores utilizados por AFIP para clasificar actividades con su correspondiente descripción. Dicha fuente, que puede obtenerse de: https://www.datos.gob.ar/fa_IR/dataset/produccion-salarios-por-departamentopartido -sector-actividad/archivo/proc 750e-4298-93d1-55fe776ed6d4.

Para realizar el análisis, se debieron hacer tareas diversas tales como la descarga de de los datos, la identificación de dependencias funcionales entre atributos, la normalización de las tablas, la construcción de un modelo conceptual de los datos y la limpieza de los conjuntos de datos con el fin de arreglar problemas de inconsistencias, datos duplicados, datos faltantes y formatos incorrectos. Además se utilizaron consultas SQL y herramientas de visualización para responder a preguntas específicas y presentar los resultados obtenidos. A lo largo de este informe, se detallarán las etapas llevadas a cabo, los desafíos enfrentados, las decisiones tomadas y las soluciones implementadas. Asimismo, se mostrarán los resultados obtenidos y se discutirá su significado en relación con la problemática planteada.

3 Decisiones tomadas

En esta sección, se presentarán las decisiones clave que fueron tomadas durante el proceso de análisis y procesamiento de los datos.

- Determinación del clasificador de actividades económicas (clae2) correspondiente a cada establecimiento: ante la ausencia de información sobre la actividad principal de los establecimientos (información necesaria para determinar correctamente la clae2 de un establecimiento), se tomó la decisión de determinar la clae2 en función del rubro y categoría de operador orgánico a la que pertenecen, y registrar esta relación en una nueva tabla. En la sección *Procesamiento de datos* se darán detalles sobre cómo se realizó este proceso.

4 Procesamiento de datos

En esta sección mostraremos las transformaciones que se aplicaron sobre los datos proporcionados, con el objetivo de evitar posibles inconsistencias, redundancias y mejorar su calidad. Dividiremos esta sección en cuatro subsecciones: *Normalización*, *Construcción de una nueva tabla*, *Diagrama Entidad-Relación* y *Curado de datos*.

Nota: Por razones de consistencia y practicidad, tanto aquí como en nuestro código hemos modificado levemente los nombres de las columnas que originalmente se denominan **Certificadora_id** y **razón social**.

4.1 Normalización

Se observó que las tablas descargadas de las fuentes ya mencionadas, no cumplían con los principios de la segunda y tercera forma normal. Esto puede llevar a anomalías de actualización de datos, como inserciones, eliminaciones y modificaciones inconsistentes, así como redundancias innecesarias. Es por esto que se decidió aplicar un proceso de descomposición de las tablas dadas con el objetivo de trabajar con tablas que cumplieran la primera, segunda y tercera forma normal, evitando así potenciales inconvenientes. En esta sección se mostrará cómo se descompusieron las tablas originales, para obtener nuevas tablas normalizadas en 1FN, 2FN y 3FN.

4.1.1 Primera forma normal

En la tabla Padrón de Operadores Orgánicos Certificados, se observó que la columna **rubro** contiene algunos valores que se podrían interpretar como que corresponden a una enumeración de subrubros. La columna **establecimiento** contiene valores que podrían interpretarse como conteniendo información de varios tipos (por ejemplo, el nombre actual de un establecimiento y también su nombre anterior, denotado por "ex"), o una enumeración de varios nombres. Incluso la columna **razon_social** podría interpretarse como que está formada por un nombre y un tipo societario (e.g., "S.R.L."). En adelante tomaremos a todas ellas como de valores atómicos ya que dentro del alcance del presente trabajo no nos interesa analizar la estructura interna de sus valores y la relación de esa descomposición con respecto a las demás columnas. Por lo dicho, a nuestros fines las tablas descargadas ya se encuentran en 1FN.

4.1.2 Tablas originales

En esta subsección expondremos las tablas originales con sus correspondientes claves primarias y dependencias funcionales.

PADRÓN DE OPERADORES ORGÁNICOS CERTIFICADOS

pais_id	pais	provincia_id	provincia	departamento	localidad	rubro	productos	categoria_id

categoria_desc	certificadora_id	certificadora_deno	razon_social	establecimiento

Dependencias funcionales:
DF_1: pais_id $\longrightarrow$ pais
DF_2: provincia_id $\longrightarrow$ provincia
DF_3: rubro $\longrightarrow$ categoria_id
DF_4: categoria_id $\longrightarrow$ categoria_desc
DF_5: certificadora_id $\longrightarrow$ certificadora_deno
DF_6: {razon_social, establecimiento} $\longrightarrow$ {certificadora_id, categoria_id}

SALARIOS DEL SECTOR PRIVADO

fecha	codigo_departamento_indec	id_provincia_indec	clae2	w_median

Dependencias funcionales:
DF_1: {fecha, codigo_departamento_indec, clae2} $\longrightarrow$ w_median
DF_2: codigo_departamento_indec $\longrightarrow$ id_provincia_indec

LISTADO DE LAS LOCALIDADES CENSALES

categoria	centroide_lat	centroide_lon	departamento_id	departamento_nombre	fuente	funcion

id	municipio_id	municipio_nombre	nombre	provincia_id	provincia_nombre

Dependencias funcionales:
DF_1: id $\longrightarrow$ {municipio_id, nombre, provincia_id, función, departamento_id, centroide_lon, centroide_lat, categoria}
DF_2: departamento_id $\longrightarrow$ departamento_nombre

DICCIONARIO DE DEPARTAMENTOS

codigo_departamento_indec	nombre_departamento_indec	id_provincia_indec	nombre_provincia_indec

Dependencias funcionales:
DF_1: codigo_departamento_indec $\longrightarrow$ {nombre_departamento_indec, id_provincia_indec}
DF_2: id_provincia_indec $\longrightarrow$ nombre_provincia_indec

DICCIONARIO DE CLASES

clae2	clae2_desc	letra	letra_desc

Dependencias funcionales:
DF_1: clae2 $\longrightarrow$ {clae2_desc, letra}
DF_2: letra $\longrightarrow$ letra_desc

4.1.3 Descomposición en 3FN

En esta subsección se mostrará cómo quedaron las tablas luego del proceso de normalización.

Notar que se incorporaron dos nuevos atributos: **id_rubro** e **id_razon_social_establecimiento**. Estos representarán identificadores únicos para los rubros y establecimientos respectivamente. Esto debió hacerse para solucionar el problema de que, algunos valores en los atributos **rubro** y **establecimiento** eran NULL y por ende, no podían ser asignados como claves primarias.

Notar también que al hacer esta descomposición en 3FN no se perdieron dependencias funcionales.

DESCOMPOSICIÓN DE PADRÓN DE OPERADORES ORGÁNICOS CERTIFICADOS

PROVINCIAS_PADRON

provincia_id	provincia

Dependencias funcionales:
DF: provincia_id ⟶ provincia

PAISES | pais_id | pais |

Dependencias funcionales:
DF: pais_id ⟶ pais

PROV_PAIS_PADRON | provincia_id | pais |

Dependencias funcionales:
DF: provincia_id ⟶ provincia

RUBROS | id_rubro | rubro |

Dependencias funcionales:
DF: id_rubro ⟶ rubro

RUBRO_CATEGORIA | id_rubro | categoria_id |

Dependencias funcionales:
DF: id_rubro ⟶ categoria_id

CATEGORIAS_ORGANICAS | categoria_id | categoria_desc |

Dependencias funcionales:
DF: categoria_id ⟶ categoria_desc

CERTIFICADORAS | certificadora_id | certificadora_deno |

Dependencias funcionales:
DF: certificadora_id ⟶ certificadora_deno

ESTABLECIMIENTOS_DATOS | id_razon_social_establecimiento | establecimiento | razon_social | certificadora_id | categoria_id | productos |

Dependencias funcionales:
DF: id_razon_social_establecimiento ⟶ establecimiento, razon_social, certificadora_id, categoria_id, productos

ESTABLECIMIENTOS | id_razon_social_establecimiento | establecimiento |

Dependencias funcionales:
DF: id_razon_social_establecimiento ⟶ establecimiento

LOCACION_ESTABLECIMIENTOS | pais_id | provincia_id | departamento | localidad | id_razon_social_establecimiento |

DESCOMPOSICIÓN DE SALARIOS DEL SECTOR PRIVADO

SALARIOS_NORM | fecha | codigo_departamento_indec | clae2 | w_median |

Dependencias funcionales:
DF: id_razon_social_establecimiento ⟶ establecimiento

DEPTO_PROV_SAL_NORM | codigo_departamento_indec | id_provincia_indec |

Dependencias funcionales:
DF: codigo_departamento_indec ⟶ id_provincia_indec

DESCOMPOSICIÓN DE LISTADO DE LAS LOCALIDADES CENSALES

LOCALIDADES_NORM | categoria | centroide_lat | centroide_lon | departamento_id | fuente |

| funcion | id | municipio_id | nombre | provincia_id |

Dependencias funcionales:
DF1: id $\longrightarrow$ {municipio_id, nombre, provincia_id, funcion, departamento_id, centroide_lon, centroide_lat, categoria}

MUNICIPIOS | municipio_id | municipio_nombre |

Dependencias funcionales:
DF: municipio_id $\longrightarrow$ municipio_nombre

PROVINCIAS_LOC | provincia_id | provincia |

Dependencias funcionales:
DF: provincia_id $\longrightarrow$ provincia

DEPARTAMENTOS_LOC | departamento_id | departamento_nombre |

Dependencias funcionales:
DF: departamento_id $\longrightarrow$ departamento_nombre

DESCOMPOSICIÓN DE DICCIONARIO DE DEPARTAMENTOS

DEPTO_PROV_INDEC | codigo_departamento_indec | nombre_departamento_indec | id_provincia_indec |

Dependencias funcionales:
DF: codigo_departamento_indec $\longrightarrow$ {nombre_departamento_indec, id_provincia_indec}

PROVINCIAS_INDEC | id_provincia_indec | nombre_provincia_indec |

Dependencias funcionales:
DF: id_provincia_indec $\longrightarrow$ nombre_provincia_indec

DESCOMPOSICIÓN DE DICCIONARIO DE CLASES

CLASES | clae2 | clae2_desc | letra |

Dependencias funcionales:
DF: clae2 $\longrightarrow$ {clae2_desc, letra}

CLAE2_LETRA | letra | letra_desc |

Dependencias funcionales:
DF: letra $\longrightarrow$ clae2_desc

4.2 Construcción de una nueva tabla

Debido a que los salarios que nos interesan analizar se encuentran relacionados con las actividades clae2 y no con los rubros y categorías de los Operadores Orgánicos, precisamos establecer una relación entre ellos. Sin embargo esta relación no surge directamente de los datos. De aquí surge la necesidad de relacionar estos atributos a través de una nueva tabla.

RCC | id_rubro | categoria_id | clae2 |

Previamente a encarar el trabajo de creación de esta tabla, se evaluó la posibilidad de sencillamente descartar los casos ambiguos, pero se consideró que eran varios los rubros ambiguos, y varios los establecimientos correspondientes a dichos rubros, con lo cual descartando se hubiera perdido información significativa.

La construcción de esta tabla se llevó a cabo mediante un proceso en parte manual utilizando una planilla de cálculo, y en parte automático utilizando Python. La parte automática del proceso puede consultarse en el

archivo **desarrollo.py**. Se tomó como base las distintas tuplas de la tabla original del Padrón de operadores orgánicos certificados, solamente considerando sus columnas **rubro**, **categoria_id** y **categoria_desc**. A ello se le agregó manualmente la columna **clae2** y su correspondiente **clae2_desc**.

Previamente a describir el proceso de asignación de clae2 a rubros, es importante notar que a cada rubro existente en la tabla le corresponde un único valor de categoria_id, sin excepción. El proceso de asignación realizado asume que vale esta propiedad.

La asignación de **clae2** se realizó en base a los posibles valores de clae2 existentes en la tabla Diccionario de clases, así como en base a la semántica del nombre de cada **rubro** y de su correspondiente **categoria_desc**.

En muchos casos, la asignación resultó obvia. En varios otros casos resultó algo trabajosa ya que la asignación podía realizarse potencialmente a más de una clae. Un ejemplo de ello es el rubro "bodega vitivinicola y elaboracion de vinagre, mermeladas, humus de lombriz" con categoría_id 2 ("Elaboradores"), al que se le podría asignar la clae2 10 ("Elaboración de productos alimenticios", por producir mermelada) o la clae2 11 ("Elaboración de bebidas", por ser una bodega vitivinícola), entre otras posibilidades. Se optó por la clae2 11 por su orden de enunciación (primero dice que es bodega y luego lo demás).

En otros casos, el nombre del rubro no era lo suficientemente descriptivo como para elegir una clae2. El rubro "elaboración", por ejemplo, figuraba en la categoría de Elaboradores y no había más información al respecto. En estos casos se desambiguó qué clae2 correspondía al rubro en base a listar los productos que los establecimientos de dicho rubro producen según consta en la propia tabla de operadores.

Para ello se agregó en la planilla otra columna llamada lista_de_productos concatenando todos los productos que producen los distintos establecimientos del rubro para el cual esta información fue tenida en cuenta durante la asignación de clae2. Esta columna sirve exclusivamente a los fines de este proceso de decisión para la asignación manual de clae2 y su documentación. En caso de duda se eligió la clae2 de manera de que ella se corresponda con la mayoría de los productos mencionados, o con el primero que fuera mencionado. En ciertos casos la decisión era aun así ambigua y se optó de manera arbitraria pero razonada en base a la descripción del rubro, su categoría y/o sus productos.

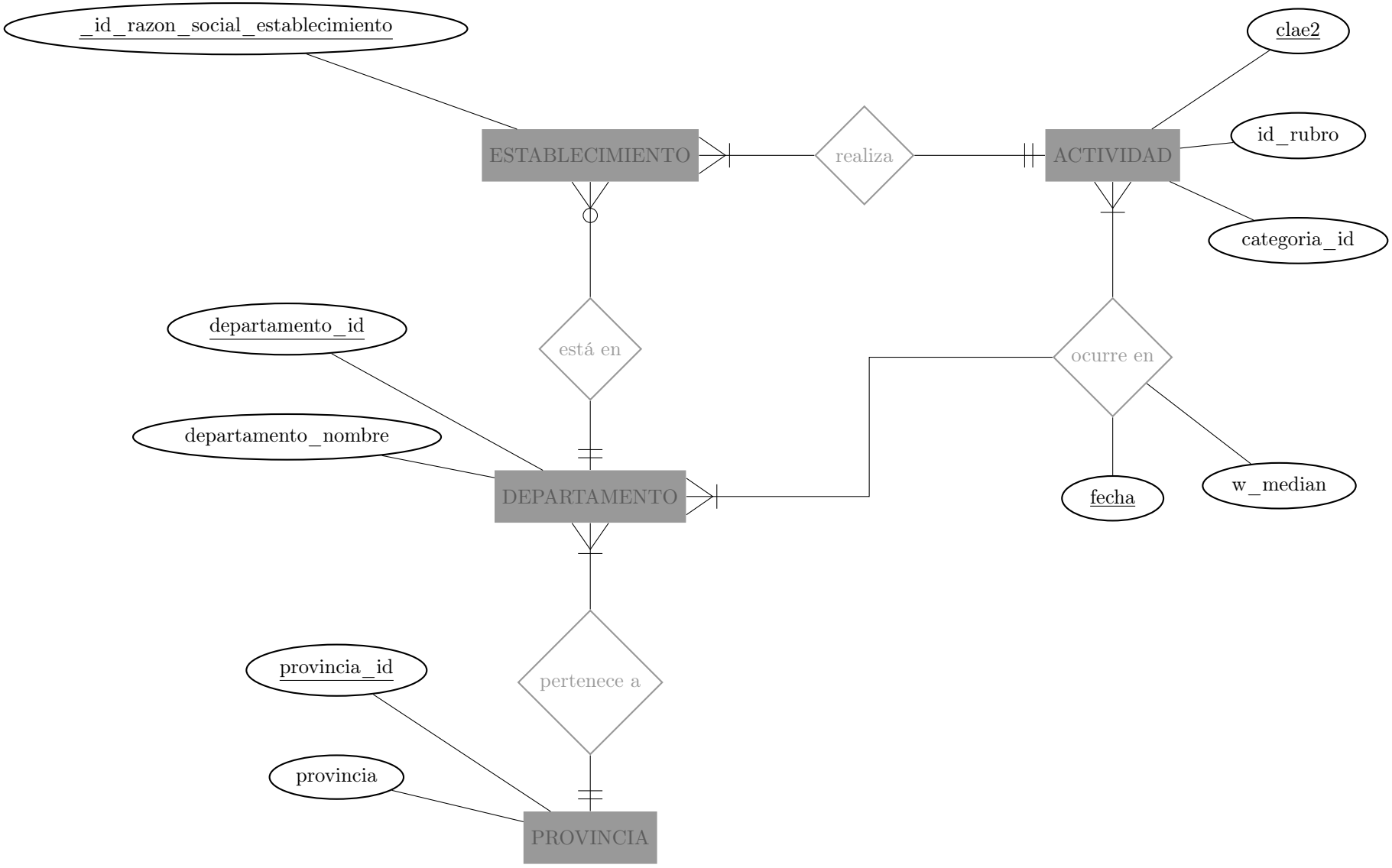
Siguiendo con el ejemplo del rubro "elaboración", se vio así que este rubro podía incluir la elaboración de productos de alimentación (aceite de soja y girasol) y de uso industrial (aceite de colza). En este caso se optó por asignarle la clae2 10 (la de productos alimenticios) por ser sus primeros productos mencionados y también por ser la clae2 correspondiente a la mayoría de los productos de este rubro en la tabla.

Otro ejemplo de falta de información suficiente es el del rubro "sin definir", del cual solo sabemos que todos sus establecimientos pertenecen a la categoría 3 ("Comercializadores"). Se presentó ambigüedad ya que no teníamos información suficiente para determinar con certeza si debían ser clasificados en la clase clae2 46 ("Comercio al por mayor excepto autos y motos") ó 47 ("Comercio al por menor excepto autos y motos"). Se consideró razonable suponer que los comercializadores minoristas de productos orgánicos, por la propia naturaleza del comercio minorista, han de tener una variedad de productos bastante mayor que la que se detalla para las empresas incluidas en este rubro, la cual sí puede ser propia de mayoristas relativamente especializados. Por ejemplo, cuesta imaginar un comercio minorista que venda únicamente yerba mate y té negro, como ocurre con uno de ellos. Por ello se optó por clasificarlos con clae2 46.

Por lo indicado, el archivo CSV resultante (TablasLimpias/rubro_categoria_clae2.csv) obra como documentación de las asignaciones manuales realizadas y de los valores que se tuvieron en cuenta para ellas, lo cual junto con el código Python en base al cual se creó la tabla logra darle trazabilidad a los cambios.

4.3 Diagrama Entidad-Relación (DER)

En esta subsección expondremos un modelo conceptual de los datos, realizado a partir de los esquemas mostrados en la subsección anterior. Para realizar este modelo conceptual se utilizó un Diagrama Entidad-Relación (DER) como herramienta.


_id_razon_social_establecimiento
ESTABLECIMIENTO
realiza
ACTIVIDAD
clae2
id_rubro
categoria_id
departamento_id
está en
ocurre en
departamento_nombre
DEPARTAMENTO
fecha
w_median
provincia_id
pertenece a
provincia
PROVINCIA

4.4 Modelo Relacional

ESTABLECIMIENTO(id_razon_social_establecimiento, id_rubro (FK a ACTIVIDAD.clae2), departamento_id (FK a DEPARTAMENTO.departamento_id))

DEPARTAMENTO(departamento_id, departamento_nombre, provincia_id (FK a PROVINCIA.provincia_id))

PROVINCIA(provincia_id, provincia)

ACTIVIDAD(clae2, id_rubro, categoria_id)

OCURRE_EN(departamento_id(FK a DEPARTAMENTO.departamento_id), clae2(FK a ACTIVIDAD.id_rubro), w_median, fecha)

4.4.1 Mapeo a tablas existentes

Por errores de organización de tareas creamos el DER en forma tardía, con lo cual desarrollamos las queries utilizando tablas que siguen la estructura de las tablas limpias normalizadas en lugar de hacerlo como hubiera correspondido, basándonos en un modelo relacional a partir del DER modelado simplificando el problema. No llegamos a tiempo a modificar las queries para que sigan el MR. Si bien hemos comprendido que ésta no es la metodología correcta ni deseable, cabe notar que se puede establecer un mapeo aproximado entre este MR y las tablas que efectivamente utilizamos en las queries, por lo cual corregir las queries para que utilicen tablas basadas en el presente modelo relacional no presentaría dificultad más allá de la falta de tiempo suficiente para ello. Mapeo:

ESTABLECIMIENTO → ESTABLECIMIENTOS_DATOS

DEPARTAMENTO → DEPTO_PROV_INDEC

PROVINCIA → PROVINCIAS_LOC

ACTIVIDAD → RCC

OCURRE_EN → SALARIOS_NORM

4.5 Curado de datos

Como ya se mencionó previamente, los datasets con los que se trabajaron tenían varios problemas, tales como, datos duplicados, presencia de valores NULL, indefinidos e inconsistencias entre las distintas tablas. Es por esto que, previo al análisis de datos se hizo un estudio de calidad de datos usando el enfoque goal question metric (GQM) y posteriormente se hizo un curado de datos para arreglar los distintos problemas que se encontraron en dicho análisis.
En esta sección se expondrá el análisis de calidad de datos realizado y se mencionarán cada una de las medidas tomadas para mejorar dicha calidad.

Goal 1: analizar presencia de tuplas repetidas en las tablas.
Question 1: dada una tabla o subconjunto de columnas de una tabla, ¿cuántas tuplas repetidas sobrantes hay?
Metric 1: porcentaje de tuplas duplicadas sobrantes con respecto a las tuplas totales útiles.

Según esta métrica, la única tabla con valores repetidos es la de Padrón de Operadores Orgánicos Certificados con un 0,5% de filas repetidos. Para reparar este problema, lo que se hizo fue eliminar las filas repetidas de dicha tabla, logrando que todas las tablas queden con un 0% de valores repetidos.

Goal 2: analizar presencia de valores NULL
Question 2: dada una tabla, ¿cuántos valores NULL hay en cada columna?

Metric 2.1: porcentaje de valores NULL con respecto a la cantidad de valores totales, para alguna columna de las dadas de una tabla.
Metric 2.2: porcentaje de valores NULL con respecto a la cantidad de valores totales, para todas las columnas dadas de una tabla.

Según la métrica 2.2 ninguna de las tablas tiene alguna fila con valores NULL en todas sus columnas.
Sin embargo, según la métrica 2.1 la tabla Padrón de Operadores Orgánicos cuenta con 0,36% de filas que contienen al menos un valor NULL en alguno de sus atributos, la tabla Salarios del sector privado un 0,28%, la tabla Listado de las localidades censales un 87,4%, la tabla Diccionario de departamentos un 0% y la tala Diccionario de clases un 1,16%. La mayoría de estos valores NULL no afectarán en el desarrollo de nuestro trabajo. Sin embargo, hay casos en los que los valores NULL podrían ser problemáticos como los que aparecen en los atributos **rubro** y **establecimiento** de la tabla Padrón de Operadores Orgánicos. Estos valores NULL dificultarían la conexión mediante JOIN con otras tablas. Para abordar este problema, se decidió crear dos nuevos atributos **id_rubro** e **id_razon_social_establecimiento** con el fin de asignar identificadores únicos a cada registro de los atributos **rubro** y **establecimiento** de la tabla Padrón de Operadores Orgánicos. Estos identificadores permitirán una conexión mediante JOIN con otras tablas incluso cuando existan valores NULL en los atributos mencionados.

Goal 3: analizar presencia de valores indefinidos ("indefinido", "indefinida", "sin definir" o "nc").
Question 3: dada una tabla, ¿cuántos valores indefinidos pero no NULLs hay en cada columna?
Metric 3: porcentaje de valores indefinidos distintos de NULL con respecto a la cantidad de valores totales, para cada columna de una tabla.

Todas las columnas de todas las tablas presentaron un 0% de valores repetidos salvo las siguientes:
- La columna **departamento** de Padrón de Operadores Orgánicos con un 0,22% de valores indefinidos.
- La columna **localidad** de Padrón de Operadores Orgánicos con un 96,54% de valores indefinidos.
- La columna **rubro** de Padrón de Operadores Orgánicos con un 7,85% de valores indefinidos.
- La columna **establecimiento** de Padrón de Operadores Orgánicos con un 29,97% de valores indefinidos.

Los valores indefinidos presentes en las columnas **departamento** y **localidad** no afectarán en el desarrollo de nuestro trabajo, por lo que se ignoraron. Los valores indefinidos de las columnas **rubro** y **establecimiento** sí podrían afectar a la hora de querer hacer JOIN entre la tabla Padrón de Operadores Orgánicos con otras tablas, pero este problema se solucionó automáticamente con la creación de los atributos **id_rubro** e **id_razon_social_establecimiento** mencionados en la métrica anterior.

Goal 4: verificar la correspondencia entre los valores únicos de dos columnas dadas de un dataframe (ej., una columna de IDs y otra de descripciones).
Question 4: dado un par de columnas ID y Descripción que se correspondan, ¿se cumple que cada ID y Descripción se corresponden unívocamente?
Metric 4: valor absoluto de la diferencia de cantidades de elementos únicos de ambas columnas.

Algunos pares de columnas dan valores mayores que cero, lo cual indicaría un posible problema, pero en los casos analizados se trataba de pares de columnas que contrariamente a lo supuesto inicialmente no debían corresponderse. Se trata de pares de nombres de localidades y sus códigos, o pares de nombres de departamentos y sus códigos. En todos estos casos la discrepancia se debe a que hay localidades y departamentos diferentes que llevan el mismo nombre, pero que lógicamente tienen asignados códigos diferentes. Esto se desambiguó manualmente mediante queries, pero también usamos unan función que, dada una lista de columnas de un dataframe, indica cuáles de las demás columnas parece depender funcionalmente de las de la lista. Por ejemplo, $fn.obtenerDFCandidata(localidades, ['departamento\_id'])$ encuentra 4 columnas que parecen, a juzgar por los datos existentes en la tabla, depender funcionalmente de la columna departamento_id. Sin embargo, $fn.obtenerDFCandidata(localidades, ['departamento\_nombre'])$ encuentra solo 1 columna. Esta discrepancia indica que ambas columnas departamento_id y departamento_nombre no dan

origen a las mismas dependencias funcionales, como sí lo harían si hubiera una correspondencia entre ellas. Nota: las DF que encuentra esta función son "candidatas", como lo indica su nombre, porque la función solo analiza los datos existentes, mientras que la DF solo existe cuando la semántica así lo indica.

Goal 5: verificar que cada columna tenga un tipo de valor único.
Question 5: para cada columna, ¿todos sus valores pertenecen a un único tipo?
Metric 5: proporción de filas correspondientes al tipo de dato más frecuente en la columna.

La mayoría de las columnas dieron un 100% en esta métrica, excepto las siguientes:
- La columna **rubro** de Padrón de Operadores Orgánicos dio tan solo un 99,64% en esta métrica.
- La columna **productos** de Padrón de Operadores Orgánicos un 99,86%.
- La columna **codigo_departamento_indec** de Salarios del sector privado un 99,72%.
- La columna **id_provincia_indec** de Salarios del sector privado un 99,72%.
- La columna **departamento_id** de Listado de las localidades censales un 99,97%.
- La columna **departamento_nombre** de Listado de las localidades censales un 99,97%.
- La columna **funcion** de Listado de las localidades censales un 14,40%.
- La columna **municipio_id** de Listado de las localidades censales un 86,3%.
- La columna **municipio_nombre** de Listado de las localidades censales un 86,3%.
- La columna **letra** de Diccionario de clases un 98,84%.

Puede observarse que esta métrica dio bastante mal para la columna **funcion** de la tabla Listado de las localidades censales, pero como este atributo nunca se utilizó en el desarrollo de este trabajo, no se hizo nada para arreglar este problema.
En el resto de los atributos que no dieron un 100% en esta métrica, por una cuestión de tiempo, se decidió también ignorar este problema, teniendo en cuenta que esto podría traducirse en una posterior pérdida de información. Por ejemplo el hecho de que solo un 86,3% de los valores de **municipio_id** sean del mismo tipo podría repercutir al responder la pregunta (ii) de la lista de preguntas a responder del trabajo práctico, donde, en un momento se debe hacer JOIN entre dos tablas, considerando justamente ese atributo.

Goal 6: evaluar y garantizar la consistencia entre ciertas columnas seleccionadas de dos dataframes.
Question 6: al contrastar las columnas seleccionadas de dos dataframes, ¿qué porcentaje de los valores resulta ser inconsistente?
Metric 6: porcentaje de valores inconsistentes entre las columnas seleccionadas de los dos dataframes.

Se encontraron inconsistencias entre las tablas localidades y deptos. Comparando sus identificadores únicos que son **departamento_id** y **codigo_departamento_indec** respectivamente, se observaron un 2,74% de inconsisntencias en las columnas correspondientes a los nombres de los departamentos y un 22,7% de valores inconsistentes entre las columnas correspondientes a los nombres de las provincias.

Goal 7: verificar consistencia de datos entre distintas tablas.
Question 7: dado un atributo A que aparece en una tabla 1 y otra tabla 2, ¿todos los valores que aparecen en el atributo A en la tabla 1, también aparecen en el atributo A de la tabla 2?
Metric 7: dado un atributo A que aparece en una tabla 1 y otra tabla 2, porcentaje de valores contenidos en el atributo A de la tabla 1 que no aparecen en el atributo A de la tabla 2.

Los resultados más relevantes fueron los siguientes:
- Los valores de la columna **departamento** contenidos en la tabla Padrón de Operadores Orgánicos contiene un 68.07% de valores que no aparecen en la columna **departamento_nombre** de la tabla Listado de las localidades censales.
- Los valores de la columna **departamento** contenidos en la tabla Padrón de Operadores Orgánicos contiene un 39.21% de valores que no aparecen en ninguna de las columnas **departamento_nombre**, **nombre** ni **municipio_nombre** de la tabla Listado de las localidades censales.
- Los valores de las columnas **departamento** y **localidad** contenidos en la tabla Padrón de Operadores

Orgánicos contienen un 33.33% de valores que no aparecen en ninguna de las columnas **departamento_nombre**, **nombre** ni **municipio_nombre** de la tabla Listado de las localidades censales.

El problema de que el 68,07% de los valores de la columna **departamento** de Padrón de Operadores Orgánicos no aparezca presente en la columna **departamento_nombre** de Listado de las localidades censales, es, particularmente grave a la hora de resolver problemas donde se involucren consultas SQL que hagan JOIN sobre estos dos atributos. Esta situación ocurre en la resolución de la pregunta (ii) del listado de preguntas a responder. Para salvar este problema, cuando se necesitó enlazar la tabla Padrón de Operadores Orgánicos a la tabla Listado de las localidades censales, de la primera, se consideró para hacer el enlace, no solo al atributo **departamento**, si no que también al atributo **localidad** y en lugar de intentar "hacer match" únicamente contra los valores del atributo **departamento_nombre**, se hizo match con los valores de las columnas **departamento_nombre**, **nombre** (de la localidad censal) y **municipio_nombre**. De esta forma quedaron solo un 33,33% de los departamentos de Padrón de Operadores Orgánicos sin poder matchearse, en lugar del 68,07% que hubiera quedado si no se hubiera tomado esta medida.

Goal 8: verificar que los valores númericos de cada atributo estén dentro del rango adecuado.
Question 8: ¿qué porcentaje de valores de una columna cuyo tipo de dato es númerico, no están dentro del rango esperado?
Metric 8: porcentaje de valores del tipo númerico que están fuera del rango esperado.

Esta métrica se aplicó únicamente a la columna **w_median** de la tabla Salarios del sector privado para verificar que solo contenga valores estrictamente positivos. La métrica mostró que un 22,42% de estos valores son no positivos. Como solución a este problema, se decidió que, cada vez que se necesitara utilizar valores de **w_median** en las consultas SQL a la hora de responder preguntas y hacer gráficos, se colocaría la condición w_median > 0 para descartar los valores no positivos.

5 Análisis de datos

Nuestro análisis de datos se ha centrado en responder las preguntas y consignas planteadas en el enunciado del trabajo práctico. En esta sección presentaremos los resultados obtenidos, destacando tanto las respuestas a las preguntas planteadas como los gráficos generados a partir de los datos.

5.1 Respuestas a preguntas

Pregunta (i): ¿Existen provincias que no presentan Operadores Orgánicos Certificados? ¿En caso de que sí, cuántas y cuáles son?

No existe ninguna provincia que no presente Operadores Orgánicos Certificados, es decir, toda provincia tiene al menos un Operador.

Pregunta (ii): ¿Existen departamentos que no presentan Operadores Orgánicos Certificados? ¿En caso de que sí, cuántos y cuáles son?

Existen 16 departamentos que no presentan Operadores Orgánicos Certificados. Estos son: CABA, Chascomús/Lezama, Ezeiza, General San Martín, Curuzú Cuatiá, Laishi, General Felipe Varela, Ángel Vicente Peñaloza, General Ortiz de Ocampo, Ñorquincó, Constitución, Juan Felipe Ibarra, Río Grande/Tolhuin, Ushuaia, Yavi y General Juan Facundo Quiroga.

Pregunta (iii): ¿Cuál es la actividad que más Operadores tiene?

La actividad económica correspondiente a la clae2 1 (Agricultura, ganadería, caza y servicios relacionados) es la actividad con el mayor número de operadores con un total de 971 Operadores.

Pregunta (iv): ¿Cuál fue el salario promedio de esa actividad en 2022? (si hay varios registros de salario, mostrar el más actual de ese año)

El salario promedio de esa actividad (clae2 1) en 2022 fue de: 128160$

Pregunta (v): ¿Cuál es el promedio anual de los salarios en Argentina y cual es su desvío?, ¿Y a nivel provincial? ¿Se les ocurre una forma de que sean comparables a lo largo de los años? ¿Necesitarían utilizar alguna fuente de datos externa secundaria? ¿Cuál?

El salario promedio anual en Argentina, considerando todas las actividades económicas, fue de 157215$ (desviación estándar: 89474$)

A nivel provincial, los salarios promedio y su correspondientes desviaciones estándar se muestran en la **tabla 1**.

Debido a la inflación de Argentina, resulta impráctico y difícil de comparar los salarios expresados en pesos a lo largo de los años. Para poder hacer esto, podría usarse una fuente de datos externa que contenga la información de la relación entre el peso y el dólar para cada mes a lo largo de los años. De esta forma, cada salario podría ser calculado en dólares lo cual facilitaría la comparación a lo largo del tiempo y mitigaría el impacto de la inflación.

Provincia	**Salario promedio ($)**	**Desvío estándar ($)**
Chubut	171588	107049
Entre Ríos	155200	95199
La Pampa	155712	105167
Mendoza	156294	114190
Río Negro	174349	106383
Catamarca	132851	100460
San Juan	140316	106504
Córdoba	150749	92365
Corrientes	143843	98911
Buenos Aires	165751	95129
Santiago Del Estero	128631	121747
La Rioja	142214	119183
Misiones	132727	98682
Santa Fe	156207	86936
Jujuy	140332	92631
Chaco	152814	161158
Neuquén	184488	106630
Santa Cruz	192892	130533
Tucuman	145247	112533
Salta	136802	81284
Formosa	132678	88325
San Luis	155401	144679
Tierra Del Fuego	193569	110138
CABA	193284	101493

Table 1: Salarios promedio y desvíos estándar para cada provincia argentina

5.2 Gráficos

En este apartado presentaremos los gráficos generados en respuesta a las consignas planteadas en el trabajo práctico y haremos un análisis sobre cada uno de ellos.

Consigna (i): Cantidad de Operadores por provincia.

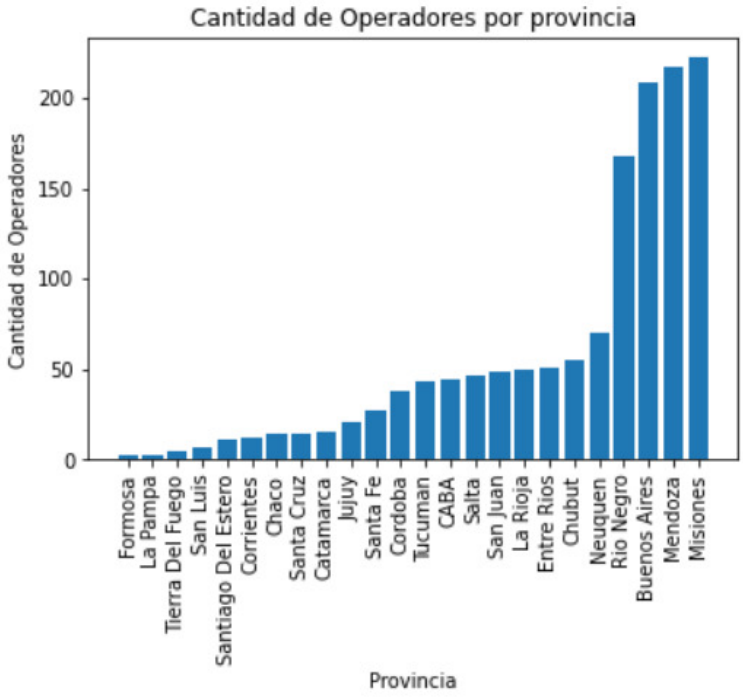


Figura 2: Cantidad de Operadores Orgánicos Certificados por cada provincia argentina

En el gráfico mostrado en la **figura 2** podemos apreciar que, las provincias Río Negro, Misiones, Mendoza y Buenos Aires destacan por tener un número de Operadores Orgánicos significativamente más alto que el resto de las provincias. Todas ellas cuentan con más de 150 Operadores. Por otro lado, las provincias que menos Operadores presentan son Formosa, La Pampa y CABA, todas ellas con menos de 10 Operadores. Como puede observarse, la variabilidad en el número de Operadores Orgánicos entre provincia y provincia, es bastante alta.

Consigna (ii): Boxplot, por cada provincia, donde se pueda observar la cantidad de productos por operador.

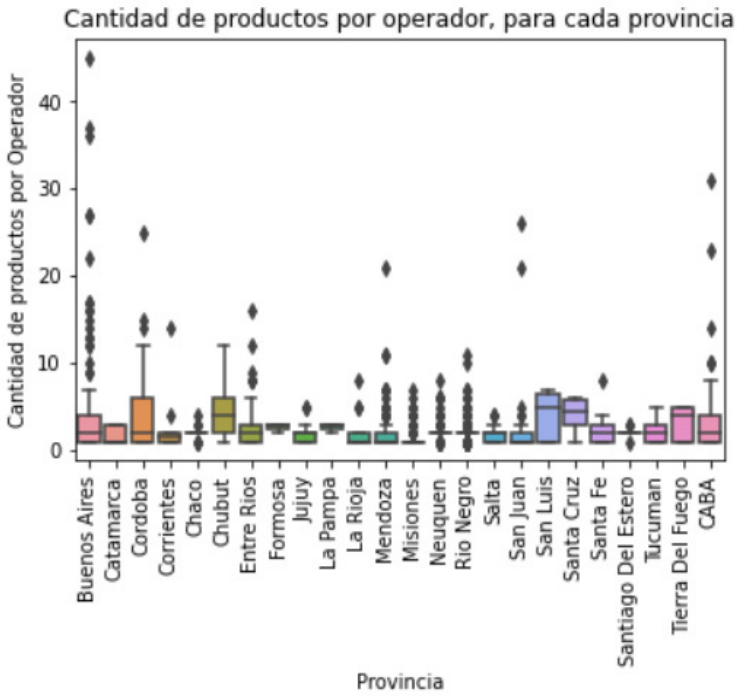


Figura 3: Boxplot que representa la cantidad de productos por Operador Orgánico, para cada provincia argentina

En la **figura 3** que ciertas provincias como Córdoba Chubut, San Luis, Tierra del Fuego y CABA tienen una gran variabilidad en la cantidad de productos que ofrecen sus Operadores Orgánicos. Por el contrario, otras provincias como Formosa, La Pampa, Misiones, Neuquén y Río Negro destacan por el hecho de que su la variabilidad es muy baja.

Consigna (iii): Relación entre cantidad de emprendimientos certificados de cada provincia y el salario promedio en dicha provincia (para la actividad) en el año 2022. En caso de existir más de un salario promedio para ese año, mostrar el último del año 2022.

Por razones de espacio, a continuación se muestra un gráfico correspondiente únicamente con Operadores Orgánicos Certificados de la clae2 1, elegida por ser la de mayor número de operadores. Los gráficos de las demás clae2 pueden obtenerse ejecutando el código del archivo **desarrollo.py**.

Para visualizar lo pedido, se decidió hacer un scatterplot de salario promedio en función de cantidad de Operadores, coloreando a cada provincia de manera diferente. El gráfico obtenido puede observarse en la **figura 4**.

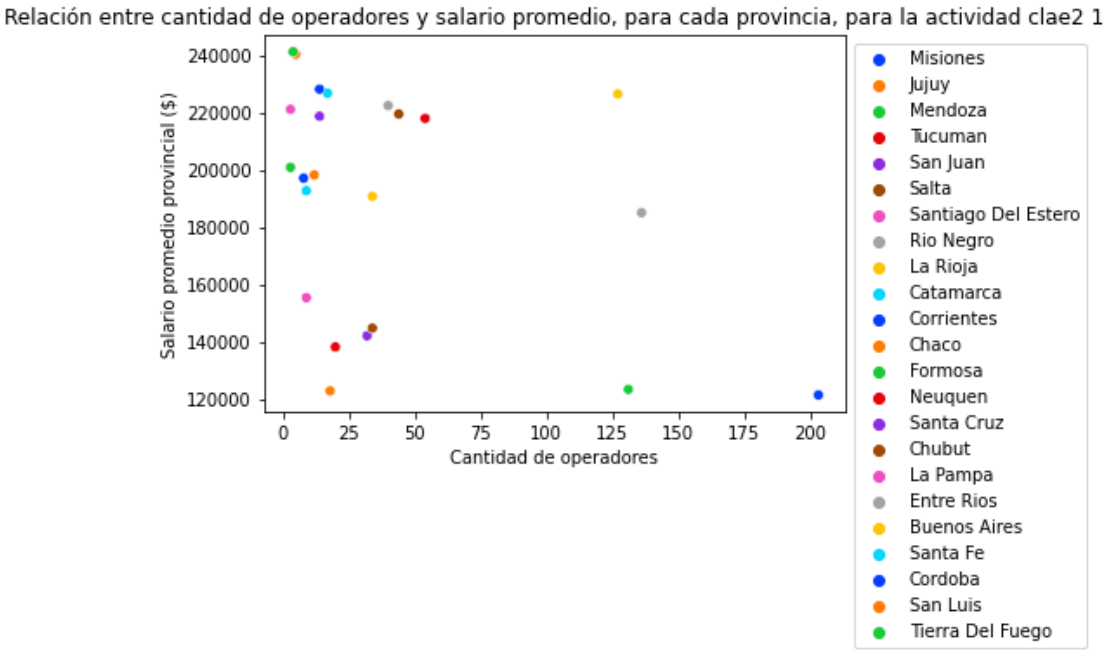


Figura 4: Salario promedio en función de la cantidad de Operadores Orgánicos, para cada provincia argentina

Consigna (iv): ¿Cuál es la distribución de los salarios promedio en Argentina? Realicen un violinplot de los salarios promedio por provincia. Grafiquen el último ingreso medio por provincia.

En la **figura 9** puede observarse el gráfico obtenido. Lo más llamativo que puede notarse de este gráfico, es que el violinplot obtenido para la provincia de Buenos Aires es significativamente más largo que el del resto de las provincias, lo cual indicaría que en esta provincia se da la mayor variabilidad de salarios promedio. Caso contrario es el de provincias como La Rioja, La Pampa, Formosa y CABA donde los violinplot obtenidos fueron particularmente cortos, lo que indicaría que la variabilidad en los salarios es baja.

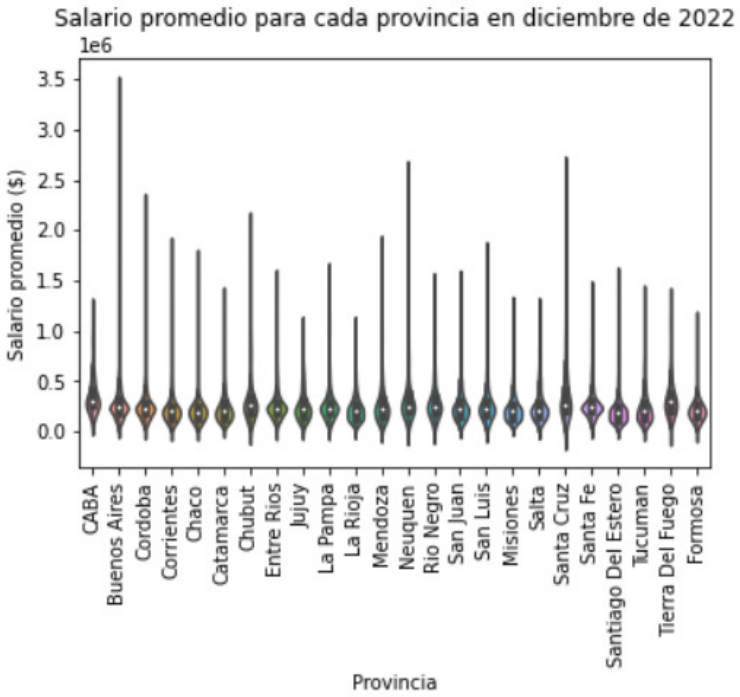


Figura 9: Violinplot que muestra la distribución de salarios promedio para cada provincia argentina.

6 Conclusiones

Para tratar de responder si existe relación entre "el desarrollo de la actividad" y el salario promedio de "los trabajadores del sector privado" en cada departamento de las provincias, conviene analizar cada parte por separado.

El salario promedio de "los trabajadores del sector privado" en cada departamento podemos calcularlo considerando todas las actividades $SALARIOS\_NORM$, sin distinguir si corresponden a operadores orgánicos pues nos interesan todas las del sector privado.

El "desarrollo de la actividad", presumiblemente haciendo referencia a la actividad de los operadores orgánicos, puede estimarse contando la cantidad de establecimientos existentes en cada departamento, y solo considerando departamentos que figuren en los datos detallados en el párrafo anterior.

La relación entre ambas cuestiones surgiría de notar una diferencia significativa entre los salarios medios de los departamentos con alto "desarrollo de la actividad", y los de bajo desarrollo, o una correlación entre ambas variables.

Al graficar esto (**Fig. 5**) encontramos que la señal que se observa es muy débil o posiblemente inexistente. En el rango bajo de desarrollo (i.e., cantidades bajas de establecimientos por departamento) se observa una gran cantidad de puntos en valores de salario medio muy distintos entre sí. Esto se debe sin duda a que en este rango la hipotética influencia del desarrollo en los salarios sería muy débil (por lo bajo del desarrollo), y a que las variadas actividades ocupan un amplio rango de salarios según muy diversas cuestiones, similar a lo que ocurre al evaluar los salarios de actividades muy diferentes.

Se observan dos outliers en el rango alto de desarrollo, lo cual achica la parte del gráfico donde se encuentra la mayoría de los puntos, que es la que nos interesa. Por ello descartamos los outliers con valor es mayores a 35, por ejemplo, y allí el gráfico va tomando más nivel de detalle. Sin embargo, la nube de puntos en el rango

inferior de desarrollo estorba. Si descartamos los valores entre 2 y 6 (**Fig. 6**), se observa más claramente una posible correlación, a la vez que se mantienen los valores con desarrollo 1 a fin de tener una idea de la escala en la que se manifiesta esa tendencia.

Finalmente podemos descartar también los valores de desarrollo 1, logrando un mejor acercamiento al área del gráfico donde la tendencia podría estar manifestándose (**Fig. 7**). Se observa así una posible correlación positiva, ya que los puntos podría considerarse como que muestran una tendencia con pendiente 5000 aproximadamente. Sin embargo, la cantidad de puntos que hemos conservado para llegar hasta esta observación resulta muy escasa, alrededor de 16, y la tendencia positiva parece depender de solo 2 de esos puntos (ambos de la provincia de Neuquén) y de otros dos más para desaparecer por completo (uno de Mendoza y otro de Entre Ríos) (**Fig. 8**).

En conclusión, parece observarse una débil señal de correlación positiva entre ambas variables, pero la escasa cantidad de datos que pudimos cruzar con ambas variables, especialmente en el rango de desarrollo cuya señal no se vería confundida por otros factores, no nos da suficiente confianza sobre si la señal es real.

Figura 5: Salario promedio por departamento vs. su cant. de establecimientos. Sin filtrar.

Figura 6: Salario promedio por departamento vs. su cant. de establecimientos. Sin outliers y filtrando [2,6].

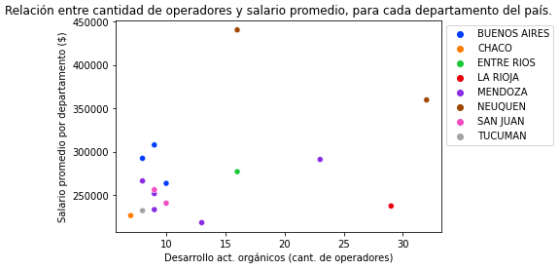


Figura 7: Salario promedio por departamento vs. su cant. de establecimientos. Sin outliers y filtrando [1,6].

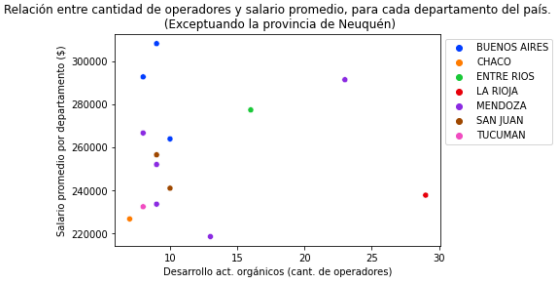


Figura 8: Salario promedio por departamento vs. su cant. de establecimientos. Sin outliers, filtrando [1,6] y sin los datos de Neuquén.