MANEJO DE DATOS Y SU VISUALIZACIÓN

LABORATORIO DE DATOS - UNIVERSIDAD DE BUENOS AIRES

Falczuk Noelia
378/22
noefalczuk@gmail.com

Fiore Juan Ignacio
259/22
juanifiore291@gmail.com

Sanes Zalazar Luna
291/22
luna.sanes@gmail.com

Fecha de entrega:
Lunes 15 de Mayo del 2023

Resumen.

Con el objetivo de colaborar con los docentes de la materia “Laboratorio de Datos” a examinar los datos obtenidos de una fuente de datos abiertos correspondientes al Padrón de Operadores Orgánicos Certificados de la República Argentina, fue necesario examinar la calidad de diseño de los mismos para poder realizar un análisis correcto y consciente.

Por esto, se ha realizado una serie de procedimientos previos de limpieza, análisis de los datos y normalización, entre otros, que incluyeron la toma de ciertas decisiones que serán documentadas, descritas y explicadas en el informe a continuación.

El desarrollo del trabajo comenzó con la descarga de datos del *Padrón de Operadores Orgánicos Certificados* y de *Salarios del sector privado*, así como del *Listado de las localidades censales según la base de datos censales del INSTITUTO NACIONAL DE ESTADÍSTICA Y CENSOS (INDEC)*, del *Diccionario de departamentos* y del *Diccionario de clases*. Luego, se continuó con su revisión y su división en nuevas tablas que respetaran la Tercera Forma Normal y tuvieran la propiedad de LOSSLESS JOIN. Cabe destacar la importancia de esta instancia en el manejo de los datos; la utilización de tablas creadas utilizando una mala calidad de diseño pueden llevar a la aparición de anomalías de actualización ~ inserción, deleción y modificación ~ así como la redundancia, aparición de valores NULL que podrían haber sido evitados y generación de tuplas espurias. También, fue importante en esta etapa la normalización de los datos para la eliminación de inconsistencias entre las fuentes de datos y dentro de ellas; las mismas serán detalladas en el cuerpo del informe.

Finalmente, a través de consultas SQL y la visualización de información mediante gráficos, fue posible ver si existe evidencia para responder a la pregunta de si hay “cierta relación entre el desarrollo de la actividad y el salario promedio que perciben los trabajadores del sector privado, en cada departamento de las provincias argentinas”.

Introducción.

Con el objetivo de buscar evidencia para responder a la pregunta de si hay “cierta relación entre el desarrollo de la actividad y el salario promedio que perciben los trabajadores del sector privado, en cada departamento de las provincias argentinas” analizamos las siguientes fuentes de datos:

- *Padrón de Operadores Orgánicos Certificados*
- *Salarios del sector privado*
- *Listado de las localidades censales según la base de datos censales del INSTITUTO NACIONAL DE ESTADÍSTICA Y CENSOS (INDEC)*
- *Diccionario de departamentos*
- *Diccionario de clases*

Sin embargo, la calidad de diseño de las mismas no era buena por lo que fue necesario realizar un procesamiento de los mismos. Para esto, se llevó a cabo la visualización de las tablas y análisis de los datos e información que contenían cada una de ellas:

padron-de-operadores-organicos-certificados.csv

La tabla “padron-de-operadores-organicos-certificados.csv” combina atributos de diversos tipos de entidades y relaciones en una misma relación. Mezcla información sobre datos propios del país, como el código y nombre; de las provincias, como id y nombre; de las certificadoras, como id y nombre, de las categorías, id y descripción; con los propios de los operadores orgánicos. También, hay columnas como “establecimiento” sin datos (NULL´s), datos en las columnas de “rubro” y “productos” que no están normalizados, que además no son atómicos (rompen la 1FN) y datos de la columna “departamento”, que mezclan nombres de departamentos con nombres de ciudades y municipios. Además, la columna “localidades” está prácticamente toda “INDEFINIDA”.

w_median_depto_priv_clae2.csv

La tabla “w_median_depto_priv_clae2.csv” contiene el salario bruto mediano de los trabajadores registrados del sector privado, por departamento/partido y clase, con frecuencia mensual y desde 2014. El problema, importante, fue la falta de datos de salarios. En la columna “w_median”, que se suponía aportaba información sobre el salario medio para cada departamento, aparecen datos de la siguiente forma “-99”. Esto indicaría que no existe o no hay información sobre salario para estos casos. Otro problema fue que dos de las filas del archivo se encontraban mal cargadas y los datos cargados en las columnas no coincidían con el tipo de datos que deberían llevar esas columnas.

diccionario_clae2.csv

La tabla “diccionario_clae2.csv” contiene los nomencladores utilizados por AFIP para clasificar actividades con su correspondiente descripción y permite asociar a la fuente primaria “Salario del sector privado” los datos de actividades. Sin embargo esta tabla contaba con algunos problemas; mezcla los datos de las actividades con las descripciones posibles para catalogar a cada una de estas actividades, lo que podría generar anomalías de actualización. También poseía en la columna “letra” una fila vacía.

diccionario_cod_depto.csv

La tabla “diccionario_cod_depto.csv” contiene los códigos utilizados por el INDEC para caracterizar los departamentos/partidos y provincias, con su correspondiente descripción, pero en el caso del código de CABA es un un código ficticio. Permite asociar a la fuente primaria “Salario del sector privado” los datos de departamento. Como la anterior, esta también poseía algunas cuestiones que tuvimos que solucionar:

En primer lugar había un error de ortografía en la escritura del nombre de la ciudad “ushuaia” que aparecía como “usuaia”. Luego, al comparar esta fuente de datos y la fuente primaria, “Salario del sector privado”, se pudo observar que a la tabla “diccionario_cod_depto.csv” le faltaban dos departamentos que sí aparecían en la fuente primaria. Finalmente, existen inconsistencias entre esta tabla y la tabla primaria “w_median_depto_priv_clae2”. Por ejemplo, Ushuaia en esta tiene el código 94014 y en la primaria, 94015.

localidades-censales.csv

El mencionado .csv permite asociar a la fuente primaria “Padrón de Operadores Orgánicos Certificados” con los datos de departamento. Lamentablemente, esta fuente posee problemas de: escritura, la ciudad “Viedma” aparecía escrita como “Biedma”; mezcla de datos; esta tabla poseía datos específicos de las localidades como latitud y longitud, nombre, su id, código y categoría, con datos de las provincias, nombre e id, de los departamentos, nombre e id, y de los municipios, nombre e id. Podría traer anomalías de actualización; y falta de datos: la columna “funcion”, “municipio _id” y “municipio_nombre” poseen muchas filas con valores NULL´s. Esto podría llegar a ser un problema en el caso que necesitamos estos datos para realizar algún análisis sobre ellos.

Además de los problemas declarados anteriormente, existían inconsistencias entre las tablas. Para luego poder trabajar con ellas en forma simultánea fue necesario solucionarlas a través de consultas SQL. En algunas tablas los datos aparecían en mayúscula y con tildes mientras que otras estaban en minúscula y las palabras sin tildes. Había inconsistencias en cuanto a cómo se nombraba a la provincia “Tierra del Fuego”. En algunos casos aparecía como “Tierra del Fuego Antártida e Islas del Atlántico Sur” y en otros casos solamente “Tierra del Fuego”. Lo mismo sucedía con cómo se nombraba a la “Ciudad Autónoma de Buenos Aires”. En algunos casos aparecía como “Ciudad Autónoma de Buenos Aires” y en otros casos solamente “CABA”.

Frente a tantos problemas, se entiende que la calidad de diseño de los datos no es buena. Esto conlleva a un más difícil entendimiento de los datos de las relaciones y a otros problemas de almacenamiento y actualización de las tuplas. Por otra parte, un mejor esquema de agrupación de atributos permite preservar la información y minimizar la redundancia, así como reducir la cantidad de valores NULL en tuplas e impedir la generación de tuplas espurias. El procesamiento de los datos permite solucionar los problemas mencionados anteriormente. Para llevarlo a cabo recurrimos a diferentes métodos, entre los que están la modificación directa de los .csv, la modificación a través de consultas SQL y a partir de funciones de la librería “pandas” de python.

Procesamiento de datos.

El proceso de procesamiento de datos incluyó la toma de varias decisiones para normalizar los datos. El objetivo es minimizar espacio de almacenamiento a través del diseño y evitar anomalías de actualización, mejorar la semántica, evitar agrupar atributos no relacionados en una misma tabla, ya que pueden generar múltiples NULL's, y armar tablas que no generen tuplas espurias.

Para lograr el objetivo es necesario colocarlas, por lo menos, en Tercera Forma Normal o 3FN, y para esto, seguir una serie de pasos:

1. **Identificación de la Clave Primaria de las relaciones (PK)**

2. **Colocar las tablas en 1FN**

La 1FN prohíbe relaciones dentro de relaciones o relaciones como valores de atributos dentro de tuplas. El dominio de un atributo debe incluir sólo valores atómicos (simples e indivisibles). Como fue mencionado anteriormente, todas las tablas se encontraban en 1FN a excepción de padrón. La tabla padrón poseía atributos que no tenían todos sus valores atómicos.

Para adaptarla decidimos, a través de consultas en SQL, dividir estos atributos en atributos atómicos. Para poder hacerlo, fue necesario crear dos nuevas relaciones: "produce" y "establecimiento_rubro". La primera relaciona la PK de la tabla con los productos que produce y la segunda relaciona la PK con el rubro al que pertenece. La consulta SQL a partir de la cual logramos colocarla en 1FN se encuentra en la sección de procesamiento de datos para la tabla "padrón" del archivo "código.py", en puntos 7, 8, 9 y 10.

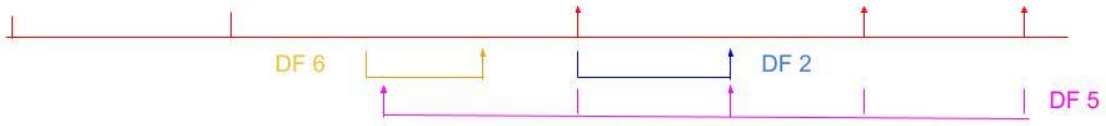
3. **Identificación de dependencias funcionales:**

Las siguientes tablas muestran las dependencias funcionales de cada una de las tablas originales (guardadas en TablasOriginales), además de, subrayadas, las PK.

En subrayado gris, la descomposición en 1FN. Las tablas "producto"y "rubros" contiene una columna con tantas filas como productos/rubros haya. Además, en "produce" y "establecimiento_rubro" tenemos una fila por cada producto/rubro distinto que cada razón y establecimiento producen.

PADRÓN

razon_social | establecimiento | pais_id | pais | provincia_id | provincia | departamento | localidad |



rubro | **productos** |categoria_id | categoria_desc | certificadora_id | certificadora_deno

Descomposicion en 3FN :

"Padron2 ": establecimiento | razon | departamento | certificadora_id | localidad | provincia_id | categoria_id

"departamento " : localidad | provincia_id | departamento | pais_id | provincia

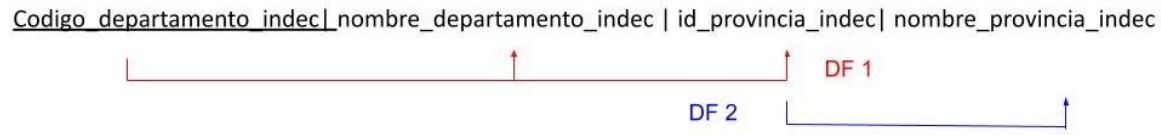
"Certificadoras": certificadora_id | certificadora_deno "paises": pais_id | pais

"Categorias": categoria_id | categoria_desc "provincias": provincia_id | provincia

"Establecimiento_rubro": razon| establecimiento | rubros "Rubros" : rubros

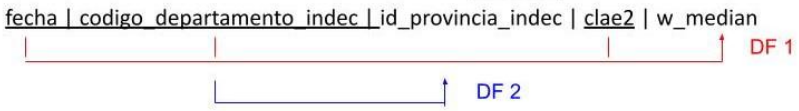
"produce": razon| establecimiento | productos "producto" : producto

DICCIONARIO DEPARTAMENTOS



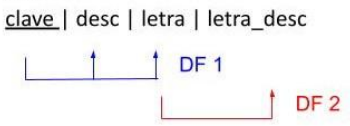
Descomposición en 3FN
"Departamentos": codigo_departamento_indec | nombre_departamento_indec | id_provincia_indec
"Provincias" id_provincia_indec | nombre_provincia_indec

SALARIOS



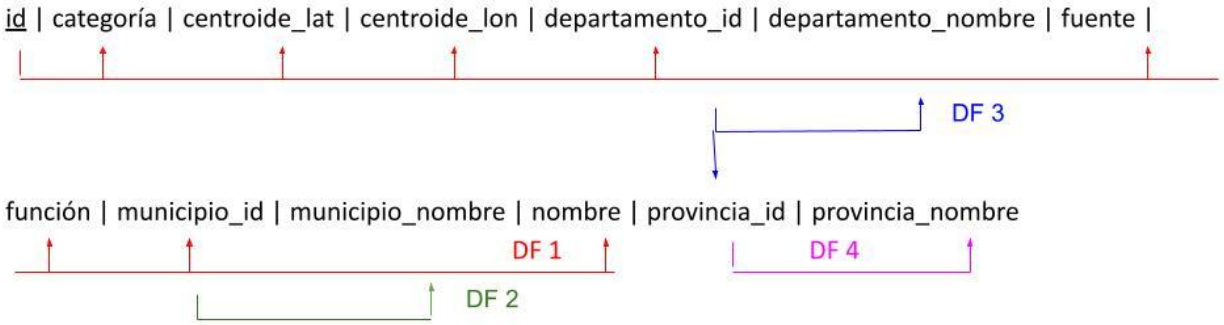
Descomposición en 3FN
"Salarios" : fecha | codigo_departamento_indec | clae2 | w_median
"Departamento" : codigo_departamento_indec | id_provincia_indec

CLAE2



Descomposición en 3FN
"Clave" : clave | desc | letra
"Letra": letra | letra_desc

LOCALIDADES CENSALES



Descomposicion en 3FN :
"localidades" : id | municipio_id | centroide_lat | centroide_lon | categoria | función | fuente | nombre
"departamentos" : departamento_id | departamento_nombre | provincia_id
"municipios" : municipio_id | municipio_nombre | departamento_id
"provincias" : provincia_id | provincia_nombre

4. **Descomposición de los esquemas en esquemas que se encuentren en 3FN y que cumplan la propiedad de lossless join.**

CLAE2

{clave} → {desc, letra}

{letra} → {letra_desc}

Entonces, se pierde una dependencia transitiva, donde: {clave} → {letra_desc} en el armado de las tablas en tercera forma normal.

SALARIOS

{fecha, codigo_departamento_indec, clae2} → {w_median}

{codigo_departamento_indec} → {id_provincia_indec}

Entonces, no se pierde ninguna dependencia en el armado de las tablas en tercera forma normal.

DICCIONARIO DEPARTAMENTOS

{codigo_departamento_indec} → {nombre_departamento_indec, id_provincia_indec}

{departamento_id} → {departamento_nombre}

{id_provincia_indec} → {nombre_provincia_indec}

Entonces, se pierde la dependencia transitiva: {codigo_departamento_indec} → {nombre_provincia_indec} en el armado de las tablas en tercera forma normal.

LOCALIDADES CENSALES

{id} → {municipio_id, departamento_id, centroide_lat, centroide_lan, categoria , función , fuente , nombre}

{departamento_id} → {departamento nombre , provincia_id}

{provincia_id} → {provincia_nombre}

{municipio_id} → {municipio_nombre}

Entonces, se pierde la dependencia transitiva {id} → {departamento_nombre , provincia_id} en el armado de las tablas en tercera forma normal.

PADRÓN

{establecimiento , razon } → {departamento, certificadora_id, localidad, provincia_id, categoria_id}

{provincia_id} → {provincia}

{pais_id} → {pais}

{razon, establecimiento} → {rubros}

{razon, establecimiento} → {productos}

{localidad, provincia_id, departamento} → {pais_id, provincia}

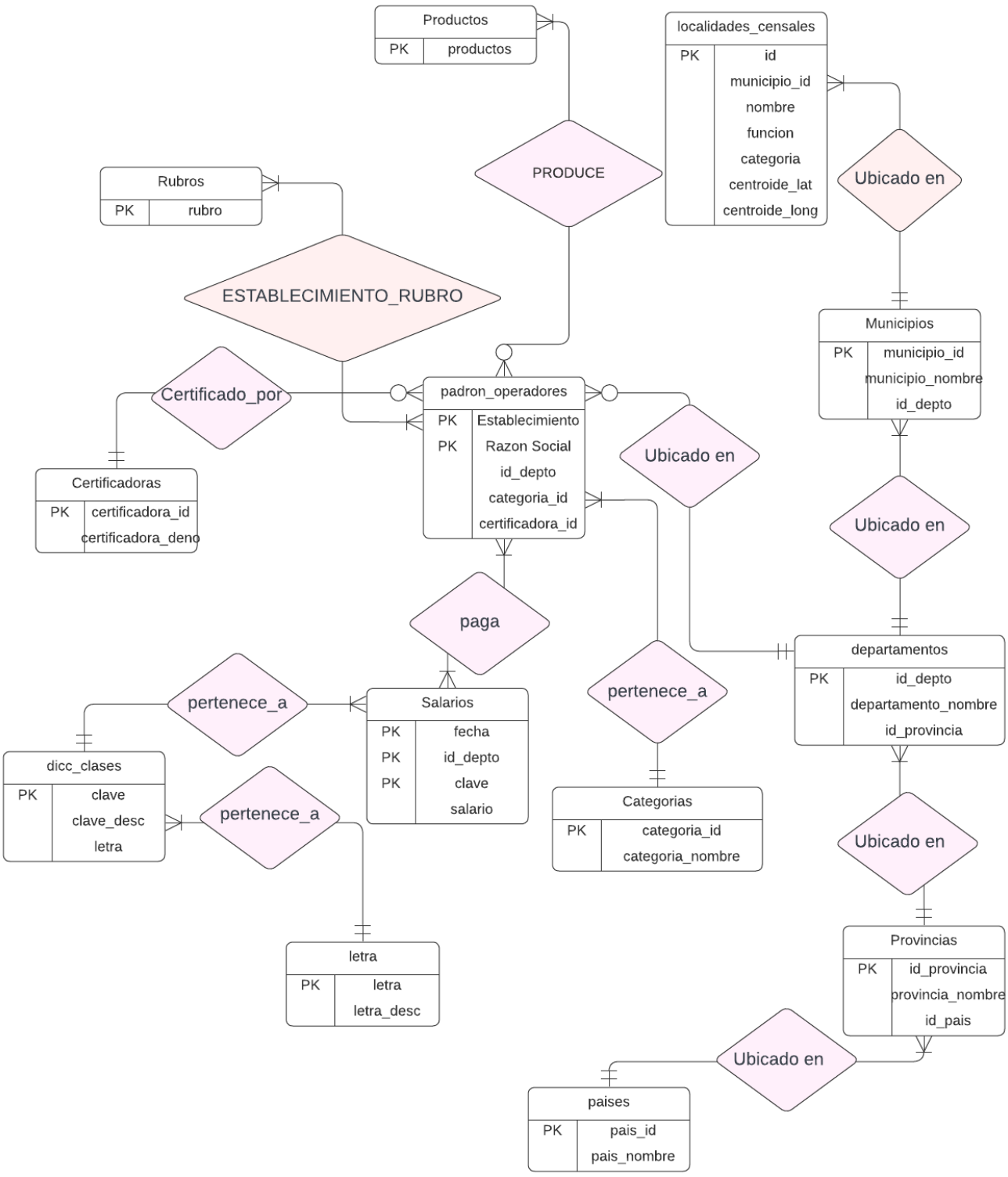
{certificadora_id} → {certificadora_deno}

{categoria_id} → {categoria_desc}

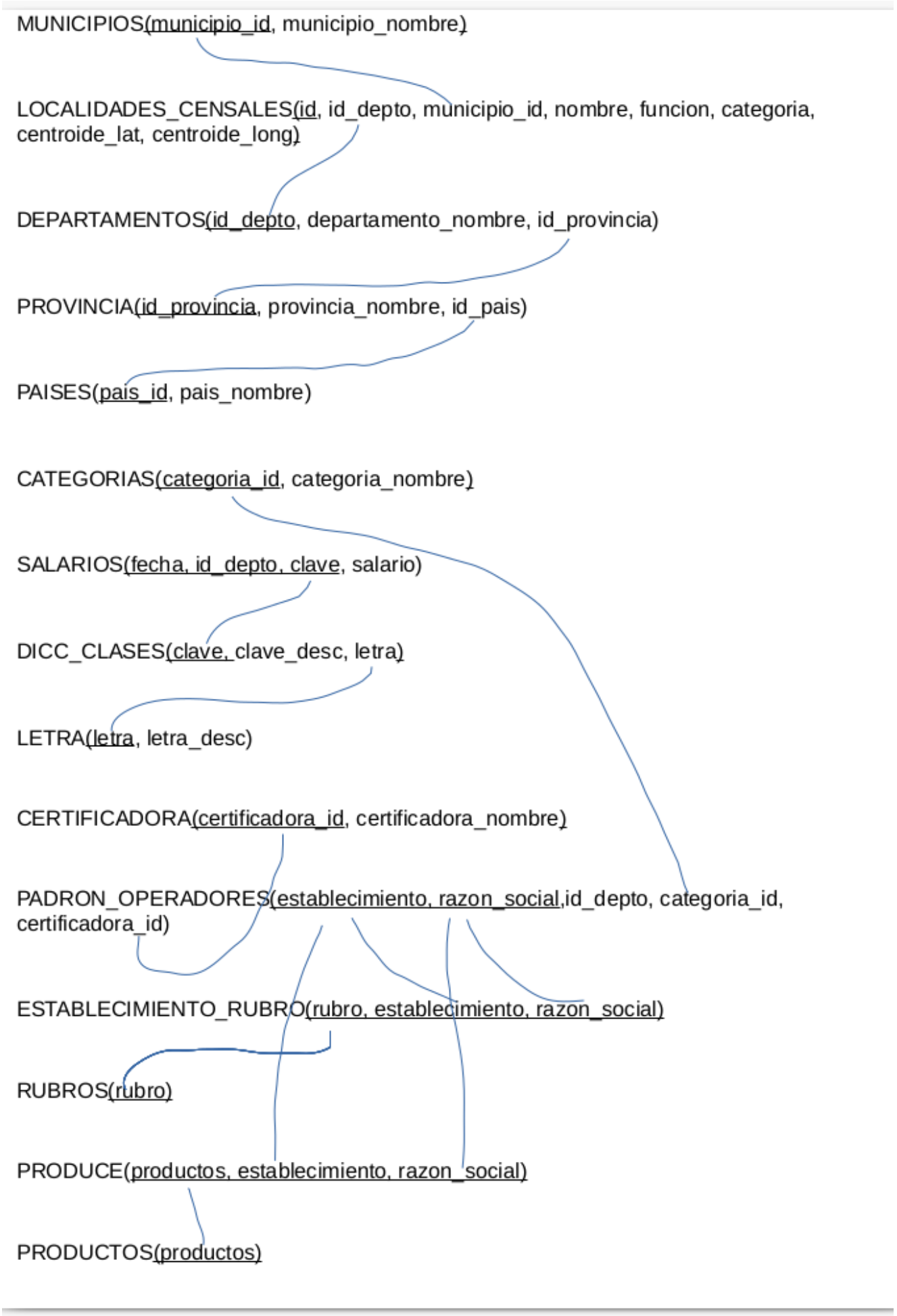
Entonces, se pierde la dependencia transitiva {establecimiento, razón} → { provincia}

5. **Construcción de un modelo conceptual de los datos.**

   Para hacer esto, utilizamos un DER como herramienta.



6. **Construcción de un modelo relacional que se encuentre en 3FN a partir del DER armado anteriormente.**


MUNICIPIOS(municipio_id, municipio_nombre)
LOCALIDADES_CENSALES(id, id_depto, municipio_id, nombre, funcion, categoria, centroide_lat, centroide_long)
DEPARTAMENTOS(id_depto, departamento_nombre, id_provincia)
PROVINCIA(id_provincia, provincia_nombre, id_pais)
PAISES(pais_id, pais_nombre)
CATEGORIAS(categoria_id, categoria_nombre)
SALARIOS(fecha, id_depto, clave, salario)
DICC_CLASES(clave, clave_desc, letra)
LETRA(letra, letra_desc)
CERTIFICADORA(certificadora_id, certificadora_nombre)
PADRON_OPERADORES(establecimiento, razon_social,id_depto, categoria_id, certificadora_id)
ESTABLECIMIENTO_RUBRO(rubro, establecimiento, razon_social)
RUBROS(rubro)
PRODUCE(productos, establecimiento, razon_social)
PRODUCTOS(productos)

Las flechas azules representan las FOREIGN KEY

7. **Corrección de los problemas de calidad de datos.**

En general, se tomó la decisión de colocar todos los valores de todos los atributos en minúsculas y sin tildes para evitar las inconsistencias mencionadas anteriormente en la introducción. También, se decidió estandarizar algunos datos: cada vez que se hacía referencia a “CABA” se pidió que se modificara este valor por “Ciudad autónoma de buenos aires” y cada vez que se nombrara a “tierra del fuego” por “tierra del fuego antártida e islas del atlántico sur”

w_median_depto_priv_clae2.csv

Para solucionar los problemas anteriormente mencionados tomamos la decisión de eliminar todas aquellas filas que en el atributo “w_median” tuvieran el valor “-99”. La decisión está basada en que si la tabla proporciona información únicamente sobre los salarios, las filas que no tuvieran este valor como atributo “w_median” no aportan información. Entonces, la eliminación de las mismas no suponía la pérdida de datos (los datos ya no estaban). Esta decisión también soluciona el problema de las filas mal cargadas debido a que estas filas desaparecen una vez llevada a cabo la decisión anterior.

localidades-censales.csv

Tomamos la decisión de reemplazar los valores NULL’s por “no tiene” en la columna “función” para mejorar la calidad de datos y disminuir la cantidad de NULL 's. Sin embargo, a las columnas “municipio nombre” y “municipio_id” no les hicimos modificaciones porque no nos eran necesarias para las visualizaciones o consultas; no son relevantes para el análisis que queremos realizar.

padron-de-operadores-organicos-certificados.csv

Para solucionar el problema de los valores del atributo “departamento” en donde se han cargado nombres de ciudades o localidades en vez de departamentos, primero chequeamos con una consulta de SQL cuáles eran los valores de este atributo que no coincidían con departamentos en el diccionario_departamentos. Para estos, modificamos con un diccionario, “a mano”, algunos de los valores que estaban mal cargados (1 y 2). Para el resto, cambiamos los valores en la columna 'departamento' que sean de municipios, por sus respectivos departamentos utilizando una consulta de SQL (4). Luego de corregir, eliminamos aquellas filas cuyo valor en columna de atributo “departamento” siga sin corresponderse con un departamento de diccionario_departamento (5).

Se tomó la decisión de agregar una columna de id_depto para el armado de los data frames y .csv´s. El objetivo es utilizar esta columna en vez de las columnas de nombre departamento y provincia y de id de departamento e id de provincia, y así, evitar en una misma tabla mezclar atributos de diferentes relaciones. Además, facilita relacionar operadores con salario y con dic_deptos (6) También, se decidió eliminar aquellas filas que en la columna de atributo “establecimiento” tuviera valores “nc”. La decisión se basó en que establecimientos es parte de la PK de padrón y debía no tener NULL 's.

Para calcular la dimensión de la calidad afectada, con ayuda de SQL, fue posible realizar un análisis del promedio de datos por provincia que se pierde. La tabla a continuación lo muestra:

id_provincia_indec	cantidadFilasSinSalarios	id_provincia_indec_2	cantidadProvTotal	PromQSeVa
6	171351	6	1009587	16
14	29159	14	189630	15
82	19238	82	156396	12
18	48855	18	156242	31
50	21618	50	139461	15
86	50667	86	135340	37
22	42956	22	130864	32
30	23640	30	127143	18
90	25190	90	119526	21
54	26838	54	115947	23
66	32094	66	114508	28
70	24282	70	110959	21
42	35219	42	101292	34
38	25406	38	83506	30
58	24444	58	82651	29
26	20803	26	78857	26
10	27024	10	78781	34
62	18065	62	72030	25
46	23655	46	60970	38
74	13245	74	55569	23
34	13850	34	47952	28
78	10117	78	47400	21
94	1082	94	17517	6

La primera y tercera columna muestran a qué provincia hace referencia, la segunda cuenta cuantas filas para esa provincia no tienen datos en salario, la cuarta cuenta la cantidad total de filas de datos para esa provincia incluyendo aquellas que no tienen datos, y la ultima, calcula el promedio de filas sin datos con respecto a las totales por provincia. En algunos casos, la pérdida de datos es muy alta porque el promedio supera el 30%. Sin embargo, la cantidad de datos es tan alta por provincia que permite hacer un análisis de los mismos.

En general, los problemas de calidad de datos en todos los datasets corresponden a problemas de procesos, existe un mal ingreso de datos como los salarios cargados con -99, que aportan filas innecesarias al dataset que no proporcionan ninguna información.

En particular, cuando hablamos de los datasets correspondientes a los códigos de departamentos y localidades censales, existe un problema relacionado con instancia, ya que muchos datos no concuerdan entre sí, como id_depto, que debería ser una manera fácil de identificar a los departamentos a través de todas las tablas y sin embargo no lo es.

8. **Generación en python los dataframes correspondientes al modelo relacional, conteniendo los datos de las fuentes primarias y secundarias.**

Una vez llevado a cabo los pasos anteriores, la construcción de los dataframes fue simple ya que, a través de consultas en SQL, dividimos en sub-tablas las originales de forma tal que cada una de ellas cumpliera con las propiedades que deben respetar para estar en 3FN.

Una de las consideraciones al crear las tablas en 3FN fue no combinar atributos de diversos tipos de entidades y relaciones en una misma relación. Por esto, la creación de las tablas en 3FN fue siguiendo el DER y el modelo relacional creado en el paso 5 y 6.

También se tomó la decisión de eliminar las columnas “localidad” de la tabla “padron2” debido a que al tener una alta proporción de valores NULL’s no aportan información suficiente para un análisis de los datos generales. Por otra parte, la eliminación de estas columnas no supone la pérdida de datos esenciales para el análisis que se quiere realizar.

Análisis de datos.

En esta sección, expondremos los resultados de las consultas de SQL, y explicaremos los gráficos resultado de los ejercicios de visualización, con el objetivo de responder a la pregunta principal del trabajo. El código correspondiente se encuentra en el archivo de Python.

Consultas SQL

1) ¿Existen provincias que no presentan Operadores Organicos Certificados? ¿En caso de que sí, cuantas y cuales son?

Si, las provincias que no presentan Operadores Orgánicos Certificados son 2, Ciudad Autónoma de Buenos Aires, y Tierra del Fuego. Esto se debe a que, con el limpiado de datos, las filas que corresponden a esas provincias se borraron, ya sea por problemas de departamento, de salario o de establecimiento.

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

Si, existen en total 354 departamentos que no presentan Operadores Orgánicos Certificados, distribuidos por toda Argentina. Los mismos se encuentran listados en la tabla departamentos_sin_op_organicos, localizada en el archivo de Python.

3) ¿Cuál es la actividad que más operadores tiene?

La actividad que más operadores tiene es Fruticultura, con un total de 854 operadores.

4) ¿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 fruticultura en 2022, en el mes de diciembre, ronda entre $309990 y $273753. Hicimos dos consultas para responder esta pregunta y ambas arrojaron resultados distintos, se pueden encontrar en el código de Python.

5) ¿Cuál es el promedio anual de los salarios en Argentina y cuál 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?

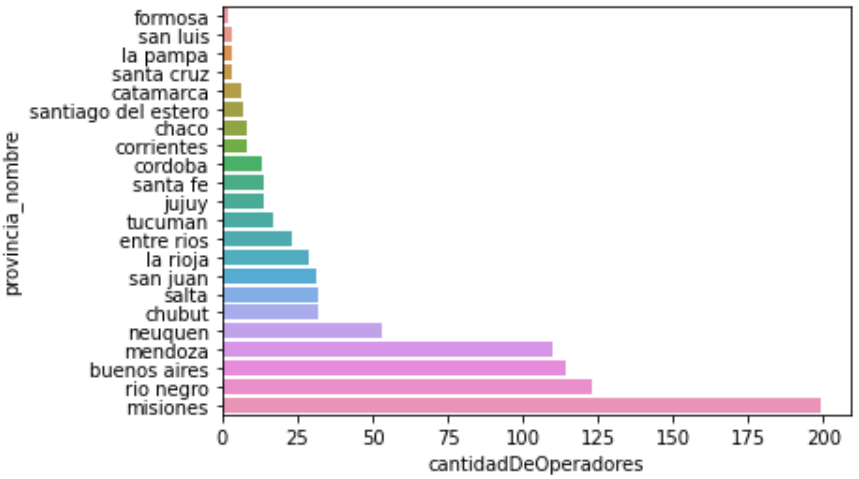
salarioPromedio	desvio	año
17139.1	12382	2016
22226.7	15877.6	2017
40686.9	31269.9	2019
55805.8	41081.1	2020
83412.1	63349.8	2021
144997	117350	2022
9847.48	7045.58	2014
12921.9	9392.77	2015
28172.5	20640.1	2018

A continuación, se exponen las tablas de promedio y desvío anual a nivel nacional. Con respecto al nivel provincial, la tabla es extensa, ya que tengo un registro por año desde 2014 a 2022, para la mayoría de las provincias del país. Puede encontrarse en el código de Python con el nombre de "PromedioDesvioProvincial". Cuantitativamente, podemos observar que el mayor salario promedio nacional se dio en 2022.

Sin embargo, es pertinente preguntarnos cómo podemos comparar de manera cualitativa los salarios de cada año, teniendo en cuenta la calidad de vida de los empleados. Para esto, nos parece que podríamos utilizar, por ejemplo, una fuente de datos que nos diga el valor de la canasta básica todos los años, y así ver si los asalariados se encuentran por encima de la línea de la pobreza.

Visualización

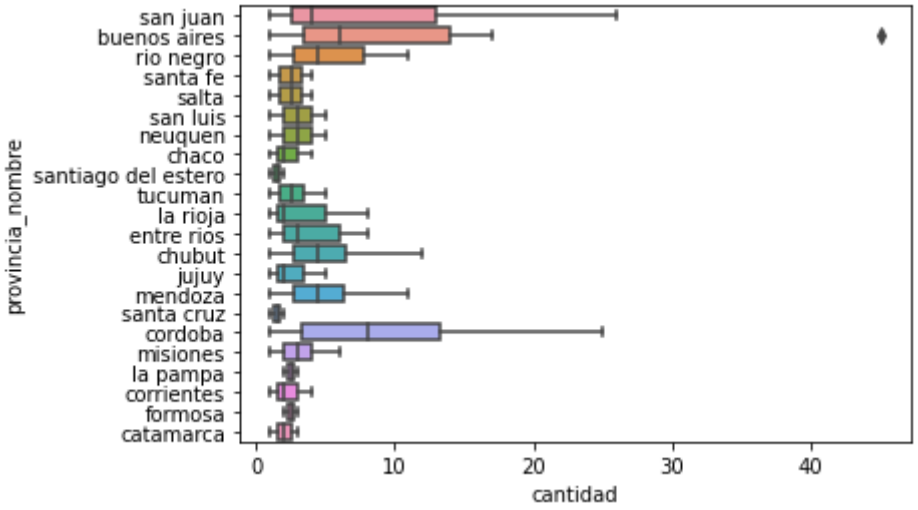
1)



El gráfico representa la cantidad de operadores (distinguidos por razón social y establecimiento) por provincia. Se observa que en Misiones hay alrededor de 200 operadores, siendo esta quien posee la mayor cantidad, mientras que Formosa, San Luis y La Pampa tienen la menor cantidad.

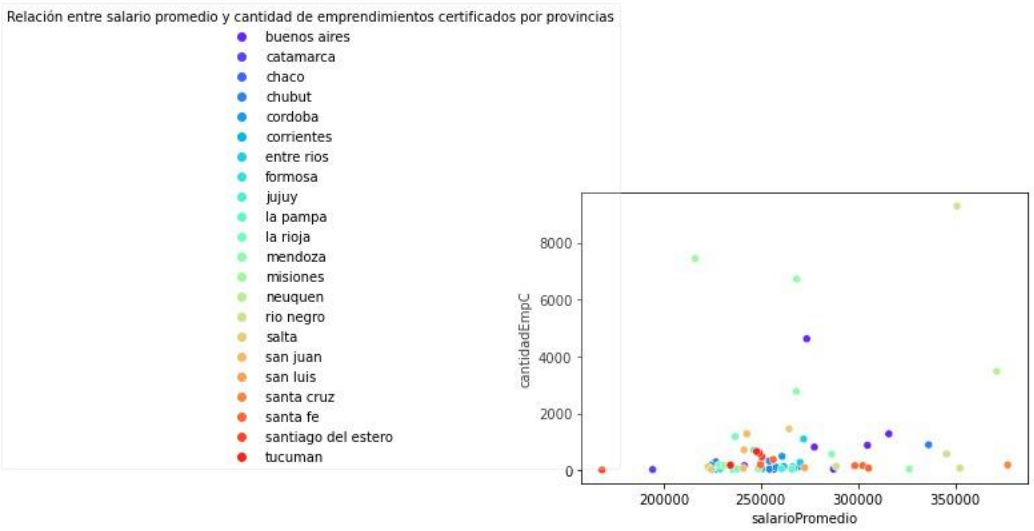
2)

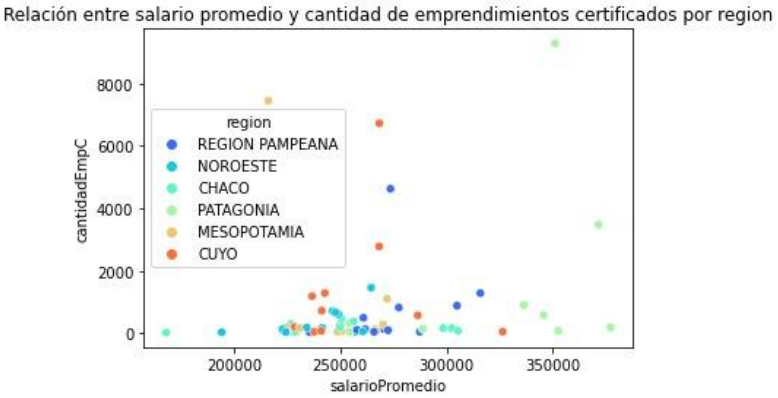
El gráfico N°2 facilita la observación de cantidad de productos por Operador, por provincia. El tipo de gráfico usado, el boxplot, representa de forma muy visual la variedad de productos. Las tapas derecha de las cajas representan el cuantil 75 mientras que la izquierda el 25, y la línea en el medio de la caja representa la mediana. Los "bigotes" del gráfico indican el resto de la distribución de datos. Así por ejemplo, tenemos que en San Juan existe una variedad amplia de productos por cada operador, siendo que los mismos tienen un rango desde 3 a 15 productos (representado por los operadores entre el cuantil 0,25 al 0,75). En Catamarca por otro lado, encontramos que los operadores ubicados allí solo tienen entre 2 y 3 productos cada uno.



3)

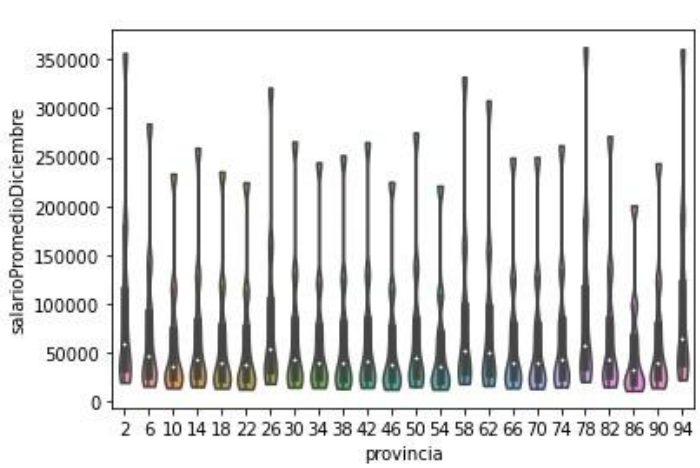
El gráfico representa la relación entre el salario promedio por actividad y la cantidad de emprendimientos que la desarrollan. Sin embargo, no podemos apreciar bien las diferencias entre tantas provincias, por lo que tomamos la decisión de agrupar por región, y así graficar el salario promedio sobre cantidad de emprendimientos para cada región. Parecería que no habría una relación entre cantidad de emprendimientos y salario promedio. No parecería existir ninguna relación linealmente dependiente (no se agrupan cerca de la diagonal con pendiente positiva) ni tampoco inversamente dependiente (pendiente negativa). Parecería que los salarios rondan principalmente entre los 230000 y 27000 independientemente de la cantidad de emprendimientos y de la región o provincia. Aunque, del segundo gráfico parecería que la región que mayores salarios tiene es la patagónica. La región del chaco es aquella que parecería tener los menores salarios.

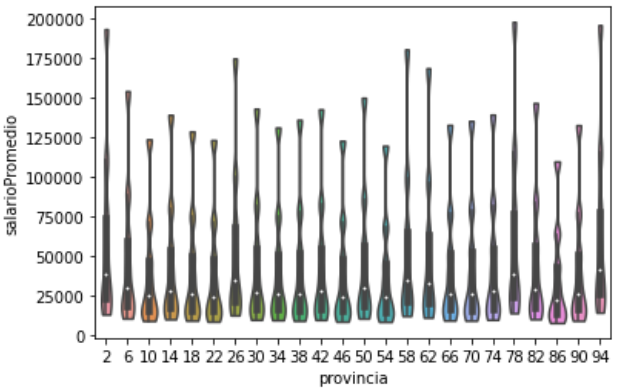




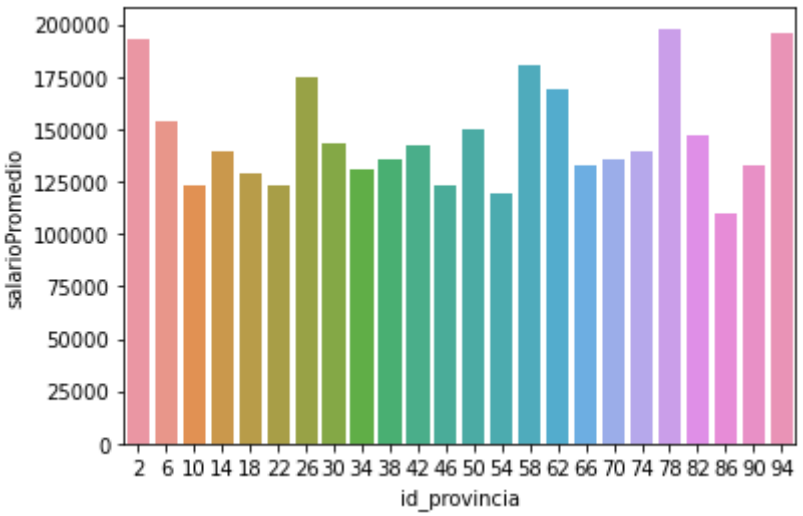
4)

El violinplot de la derecha muestra la distribución de salarios promedio por provincia. De provincia, elegimos mostrar el código ya que sino, los nombres son ilegibles en cualquiera de los dos ejes. Se pueden encontrar los nombres de las provincias en la tabla “provincias”, ubicada en la carpeta TablasLimpias. El gráfico violinplot es similar a un boxplot, con la diferencia que me muestra una estimación de la densidad de los datos de cada variable. El gráfico de la izquierda corresponde a la distribución de salarios promedio por provincia, para el mes de Diciembre.

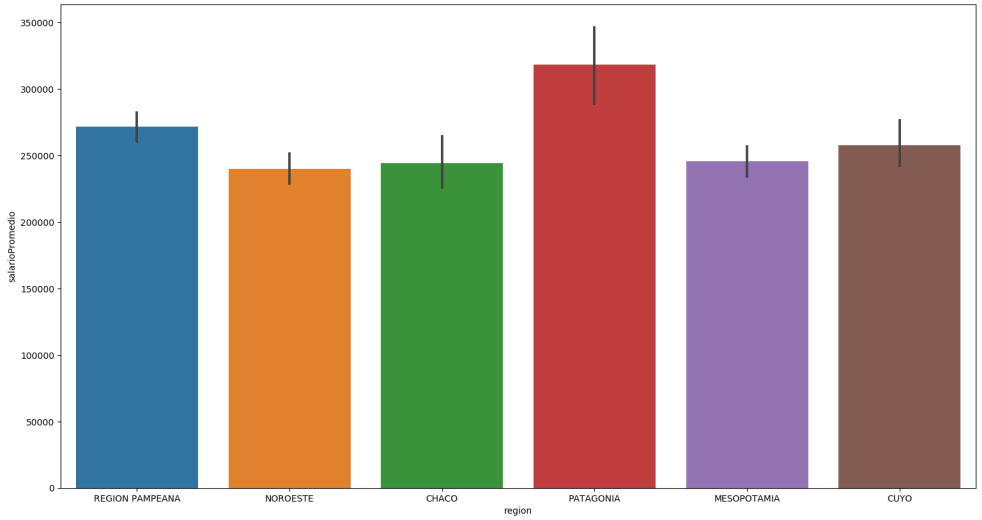




Este gráfico de barras representa el último ingreso promedio por provincia, es decir, en diciembre de 2022. Podemos observar que Tierra del Fuego (cod 94), Ciudad Autónoma de Buenos Aires (cod 2) y Santa Cruz presentan los salarios promedios más altos en este mes.

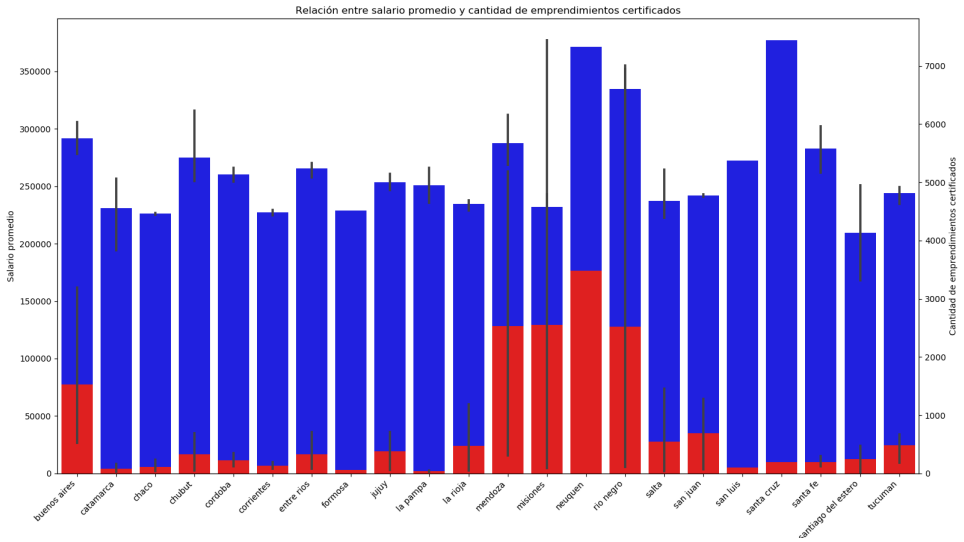


Al ver el gráfico de salario promedio por regiones, Patagonia es efectivamente el que presenta salarios más altos.

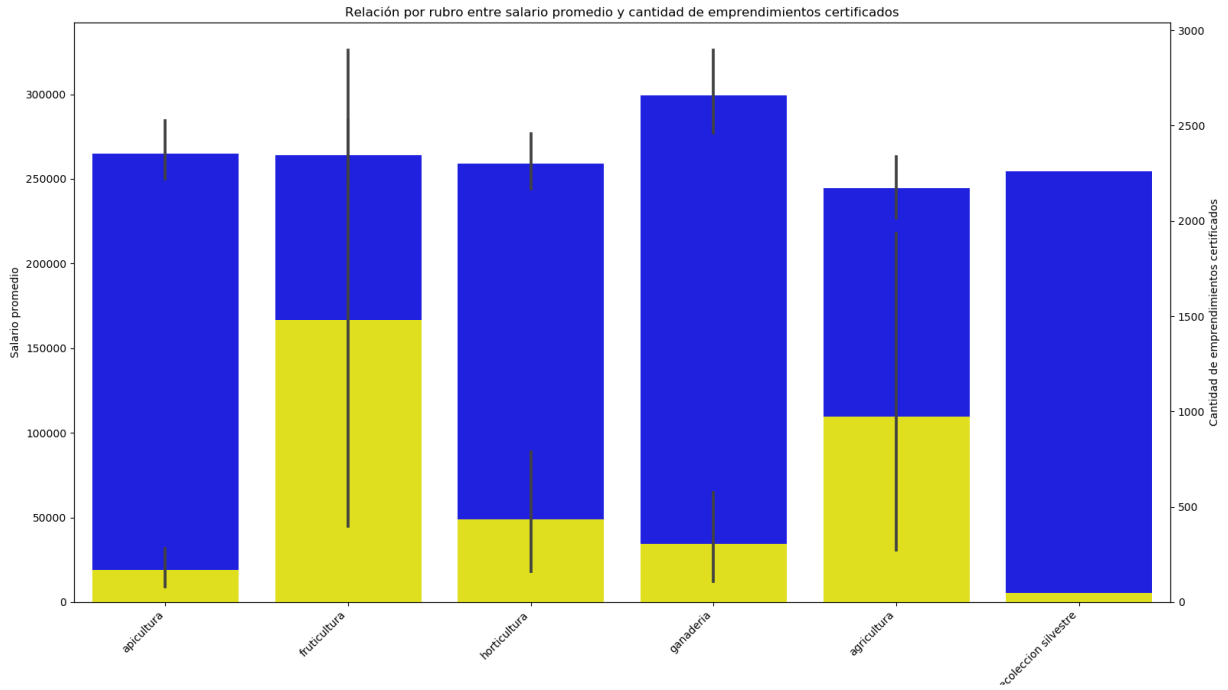


5.

Este gráfico muestra la relación del salario y la cantidad de operadores en cada provincia. (doble eje Y)



6. Este gráfico muestra la relación entre el salario promedio y cantidad de operadores por cada rubro.



Conclusiones.

A pesar de las conclusiones hechas a partir de las consultas SQL y las visualizaciones, no existe suficiente información y evidencia como para afirmar que exista una relación entre el desarrollo de las actividades en los departamentos y el salario promedio en los mismos. No solo influye la falta de datos que ofrecen las tablas, sino también las decisiones tomadas en el armado de tablas que permitieran la Tercera Forma Normal. Como fue mencionado, la eliminación de filas de la tabla padrón donde en la columna “establecimiento” existían valores NULL, y el mal armado de la columna departamento en esta misma tabla, hizo que se perdieran una gran cantidad de datos. Entre ellos, se encuentran todos los Operadores Orgánicos de Tierra del Fuego y Ciudad Autónoma de Buenos Aires.

Teniendo en cuenta esto, se puede igualmente decir que lo más probable es que no exista una relación entre el desarrollo de actividades en los diferentes departamentos y el salario promedio en los mismos. Por ejemplo, el gráfico III muestra claramente esto.

Hemos realizado dos gráficos extras para ver si existe relación entre el rubro y el salario (gráfico 6) y ver la relación entre el salario y cantidad de operadores por cada provincia específicamente (gráfico 5). Estos muestran que no necesariamente mayor cantidad de operadores implica mayores salarios, tanto por regiones como por rubros. No se encuentra patrón respecto al aumento o disminución de cada parámetro.