Trabajo Práctico 1

Desarrollo de la producción orgánica y proporción de mujeres empleadas en establecimientos productivos

Materia: Laboratorio de Datos

Grupo: How DER You?

Azul Barracchia
Augusto Guarnaccio
Mario Sigal Aguirre

Resumen

Este informe documenta todo el proceso que tuvo que ser llevado a cabo a raíz de la necesidad de responder a la pregunta de si existe cierta relación entre el desarrollo de la actividad orgánica y la proporción de mujeres empleadas en establecimientos productivos de las provincias argentinas.

En este se podrán encontrar detalladas las decisiones tomadas en el proceso de limpieza, en el armado de los data frames y en el modelo que se planteó para relacionar las fuentes. Además se encuentra un análisis de los datos a partir del trabajo de relacionar las entidades mediante consultas SQL y gráficos.

Introducción

Con este trabajo de manipulación y análisis de datos se desea conocer si existe una relación entre la proporción de mujeres empleadas en el área de producción de cada departamento con el desarrollo de la actividad de producción orgánica.

Para esta investigación la información fue tomada de diversas bases de datos del gobierno nacional. Estas son el Padrón de operadores orgánicos certificados, la Distribución geográfica de establecimientos productivos (ambas fuentes primarias de este trabajo), las Localidades de la base de asentamientos humanos de la República Argentina y el Diccionario de CLAE (ambas fuentes secundarias).

Para alcanzar dicho objetivo es necesario plantear un modelo de datos del problema. Así en primer lugar, será necesario generar un Diagrama Entidad-Relación que permita visualizar las entidades del modelo junto con sus propiedades y también entender la forma en la que estas se relacionan con el fin de poder mostrar la relación que compete a este trabajo.

En segundo lugar, se deberá realizar el pasaje del DER al modelo relacional para así poder empezar a manipular los datos. No obstante, como se desconoce la calidad de los datos de las fuentes, estas deberán ser sometidas a testeos para conocer su forma normal y la calidad de los datos.

En tercer lugar y ya con las relaciones dispuestas en tablas que satisfagan la 3FN, se responderán algunas preguntas necesarias para el análisis de los datos mediante la utilización de consultas SQL y herramientas de visualización.

Así, en las siguientes secciones del informe el lector se encontrará con la documentación de las decisiones tomadas en el transcurso de la investigación; con la descripción de los procesos que se siguieron para aumentar la calidad de los datos, conocer las formas normales de las fuentes y para la importación importar los datos; y con el análisis de datos del cual se extraerán las conclusiones finales de la investigación.

Decisiones tomadas

En esta sección se encuentra una breve descripción de decisiones importantes que por su generalidad o poca correlación con los incisos posteriores no se encuentran especificadas en estos.

En la tabla Localidades original en los valores correspondientes a departamento e id_departamento de las comunas de la Ciudad Autónoma de Buenos Aires se encuentran “enters” o “\n” antes del cierre de las comillas del string. Esto genera que no se identifiquen correctamente los valores en las comparaciones que realizan las consultas SQL. En este caso, se decidió borrar estos “enters” en ambos valores manualmente.

En el data frame generado, df_Producto, se decidió agregar los valores de la columna clae2 para poder tener una clasificación (aunque sea mínima) de cada producto. Esto se hizo de forma manual mediante la función .at() de la librería Pandas (la cual permite asignar valores en un data frame) ya que no se desconoce de algún algoritmo que pueda clasificar estos productos de manera correcta y automática.

En el armado del data frame df_Operador_organico y la asignación del atributo id_departamento para la relación, se decidió eliminar los valores de operadores orgánicos a los cuales no se les pudo asignar su id_departamento y, por lo tanto, quedaban en NULL. Los motivos de esta decisión fueron el deseo de mantener la calidad de los datos, la necesidad que id_departamento siga siendo una foreign key y la importancia de evitar la pérdida de datos a la hora de vincular esta tabla con el data frame df_Departamento. A su vez, es importante aclarar que aquellos valores descartados representan un 9% del total. Esto generó cierta preocupación pero frente a la imposibilidad de mapear correctamente los valores restantes se desistió en la tarea.

En la tabla de localidades, algunos id_departamento aparecen de forma repetida para distintos departamentos. Se decidió localizar cada uno de estos y cambiar manualmente estos id_departamento por ids nuevos y no repetidos.

En el data frame df_Establecimiento_productivo se decidió que éste sólo contenga los establecimientos productivos de índole orgánico (Clae2 ϵ {1,2,3,10,11,13,20,21}).

Por último, existe la posibilidad de que se encuentren consultas SQL repetidas. La decisión de mantener estas consultas repetidas recae en que ayuda y facilita la lectura de cada ítem sin tener que desplazarse por el largo del archivo.

Procesamiento de Datos

Formas normales

A la hora de la recopilación de los datos se recibieron tablas poco descriptivas haciendo la comprensión de estas una labor costosa posteriormente. Cuando se analizó la forma normal de cada una se pudo comprender el por qué.

La tabla del padrón de operadores orgánicos no se encuentra en 1FN. Esto se puede ver en la columna “Productos” donde en algunos casos hay más de un atributo en las filas correspondientes por lo que no hay valores atómicos e indivisibles.

La tabla de distribución de establecimientos productivos tampoco se encuentra en 1FN. Si bien es un detalle, hace que tampoco se cumpla la condición de valores atómicos.

En la columna “Empleo” se pueden encontrar valores de la forma “letra. rango de números”; valor que podría dividirse en otras dos columnas.

La tabla de localidad tampoco se encuentra en 1FN. En la columna “nombre_aglomerado ” se pueden hallar más de un atributo en algunas filas (aquellas filas donde la columna “tipo_asentamiento” encontramos el valor “componente de localidad compuesto”) por lo que no tenemos valores atómicos e indivisibles.

La tabla del CLAE está en 1FN ya que el dominio de los atributos incluye solo valores atómicos y no tiene relaciones dentro de relaciones o relaciones como valores de atributos dentro de tuplas. También está en 2FN ya que las dos claves posibles son de un solo atributo (clae6 y clae6_desc). Sin embargo, no está en 3FN. Esto puede ser demostrado mediante un contraejemplo y, si bien se expondrá uno, puede no ser el único. El contraejemplo hallado es la DF clae3 → clae3_desc. Esta DF pertenece al conjunto de DFs de la tabla pero tiene del lado izquierdo un atributo que no es SK de la relación y, por lo tanto, la tabla no está en 3FN.

Como se expuso, las relaciones con las que se va a trabajar carecen de calidad. Es por este motivo que se llevó a cabo una limpieza y un proceso de mejora de datos para así poder analizar la información no solo con mayor facilidad sino de una forma correcta.

A su vez, cabe recalcar que además de estos procesos también se siguieron una serie de pautas para asegurar que las relaciones no solo esten en 3FN sino que también cumplan con ciertos estándares de calidad. Estas pautas fueron 4.

La primera pauta es la de semántica. Esta nos dice que los esquemas diseñados deben ser fáciles de explicar y que por ello no se pueden combinar atributos de diversos tipos de entidades y relaciones en una misma relación.

La segunda pauta es la de almacenamiento. Esta nos dice que para evitar anomalías de actualización no se deben almacenar natural joins.

La tercera pauta es la del manejo de NULLs. Esta nos dice que se deben evitar asignar atributos a relaciones cuando estos puedan ser frecuentemente NULLs.

La cuarta pauta es la de tuplas espúreas. Esta nos dice que se deben evitar a toda costa relaciones que generen tuplas espúreas.

Mejoras en la calidad de datos

Un análisis de la calidad de los datos y la mejora de estos es importante antes de cualquier trabajo con bases de datos. En las cuatro fuentes de datos se encontraron inconsistencias, repeticiones, valores NULL, y datos no relevantes para nuestro análisis. Se realizaron tareas de mejora de datos y se hizo una limpieza para que queden solamente valores relevantes.

En la fuente de datos de Padrón de Operadores Orgánicos Certificados, se hizo evidente que no existía ninguna clave ya que todos los atributos tenían repetidos. Por esto, se tomó la decisión de agregar un id que funcione como clave. Además, se encontraron valores NULL en la columna establecimiento descriptos como “NC”. Este problema está asociado al desconocimiento de los datos durante la instancia de recolección. Para mantener la completitud del atributo establecimiento y procurar un mejor entendimiento, se reemplazaron los valores “NC” por “ESTABLECIMIENTO ÚNICO”. Esto se decidió así ya que los operadores orgánicos cuyo establecimiento era desconocido tenían distintas razones sociales. Esto significa, entre otras cosas, que son todos entes distintos, y en sí, que cada uno tiene un solo establecimiento.

Además, se decidió desglozar las listas de productos de cada operador. Para esto, primero se reemplazaron algunos caracteres no deseados con el afán de mantener los productos en listas de valores separados por comas, a las cuales posteriormente se les aplicó la función UNNEST (función que desglosa las listas en valores atómicos) y luego la función TRIM que saca los espacios al principio y final del string. Por último, se eliminaron otros caracteres no deseados de la columna Producto, como “)”, “.”, “(”. Al finalizar la limpieza, quedó una tabla consistente con seis columnas: id, establecimiento, razón social, departamento, id_provincia y producto.

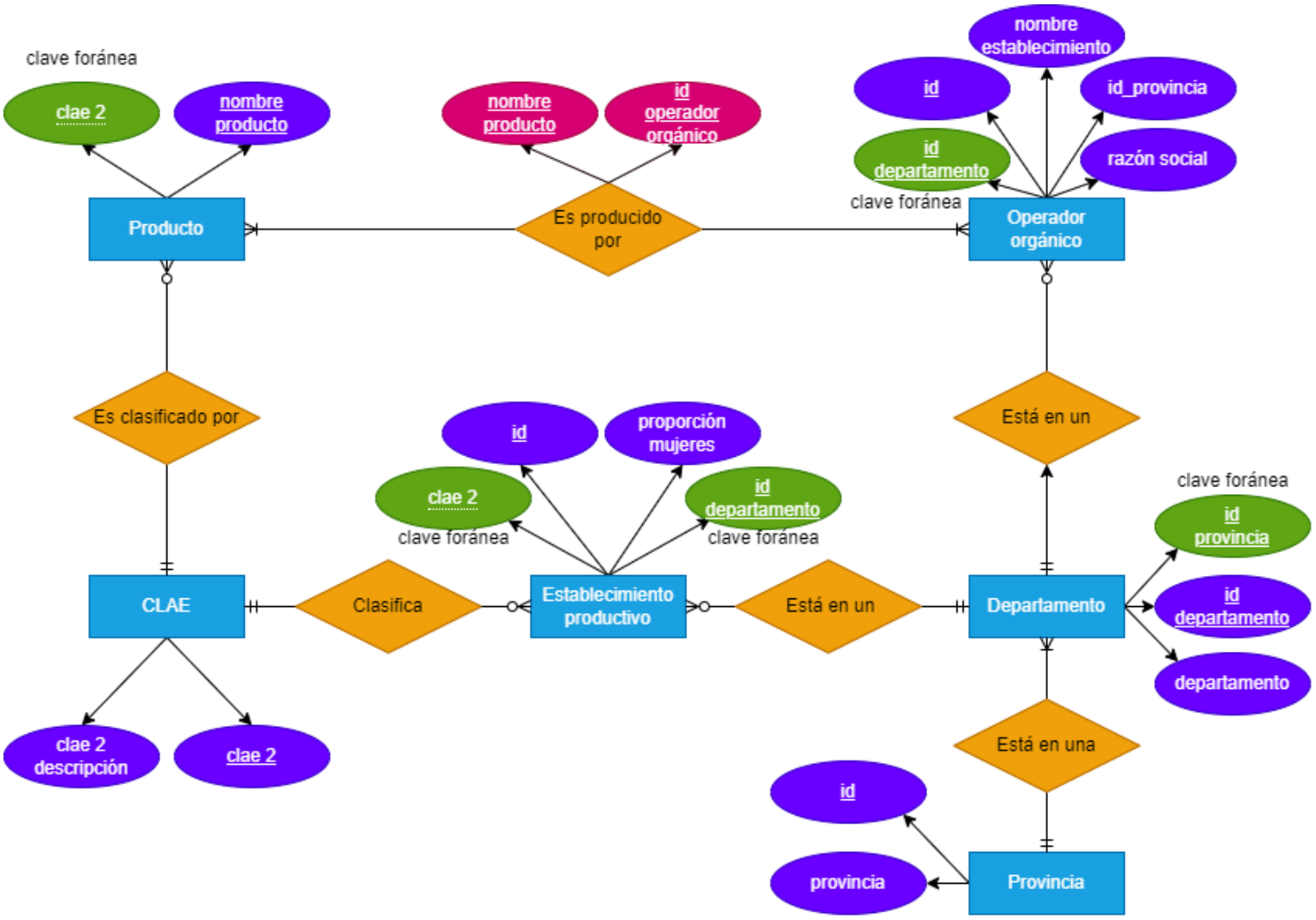
En la base de datos de Localidades, se cambió a mano los valores de Id_departamento y Departamento para Ciudad Autónoma de Buenos Aires, ya que no eran atómicos, y se reemplazaron por un solo valor. En esta tabla, luego de una limpieza de las columnas no deseadas, se hizo evidente la aparición de filas repetidas. Para evaluar la magnitud de este problema buscamos la cantidad de filas repetidas en relación al total. Aproximadamente, un 85% de las filas eran repetidas, se arregló este problema con un SELECT DISTINCT. También se eliminaron las tildes de Departamento, para facilitar las búsquedas y tener un criterio unificado. Luego de examinar la tabla con cuidado, se encontró que había ids de departamento repetidos para distintos departamentos, así que se cambiaron estos pocos valores a mano. Finalmente, Localidades quedó con cinco columnas: id_departamento, id_provincia, provincia, asentamiento y departamento.

La tabla de Clae también tenía datos irrelevantes en relación a la investigación. Por ese motivo, se hizo una limpieza para quedarse solo con la columna de clae2 y su descripción. Además, se filtró por las descripciones de clae2 para que queden sólo las relevantes y así achicar la tabla significativamente, para simplificar su lectura.

Para la tabla de Establecimientos Productivos se decidió quedarse con parte de sus datos. Para analizar el atributo de relevancia de los datos, se formuló la siguiente pregunta: ¿cuántos establecimientos productivos son de productores orgánicos?

Para responder esta pregunta se recurrió a los ids de clae2 que corresponden a productos relacionados a la producción orgánica. El resultado fue que el 13 % de los establecimientos eran relevantes para la investigación. Para solucionar esto, se realizó una subquery para que quede solo ese 13% en la base de datos. A su vez, esta fuente de datos tenía varias inconsistencias en los valores que almacenaba en id_departamento en comparación con la tabla de localidades. Por ejemplo, había ids menores a 10000 (en la tabla localidades no sucede) como también los ids asociados a CABA eran distintos. Por ello, se decidió cambiar los ids de departamento para que tengan el mismo formato y puedan compararse. De esta forma, quedó una tabla limpia, con las columnas: id, id_departamento, clae2 y proporcion_mujeres.

Diagrama entidad relación



Pasaje al modelo relacional

El pasaje del DER al modelo relacional genera las siguientes relaciones, todas en 3FN (luego cuando se planteen las DFs se justificará esto). Primero, serán enumeradas las tablas que hagan referencia a entidades del DER y luego las que hagan referencia a relaciones muchos a muchos.

Producto	
producto	clae2

Clae	
clae2	clae2_descripción

Establecimiento_productivo			
id	proporción_mujeres	clae2	id_departamento

Provincia	
id	provincia

Departamento		
id	departamento	id_provincia

Operador_orgánico				
id	establecimiento	id_provincia	razón_social	id_departamento

Es_producido_por	
producto	id_op_or

En el siguiente cuadro se pueden ver la PK, la FK y las DFs de cada una de las tablas.

Relación	PK	FK	DFs
Producto	producto	clae2	producto → clae2
Clae	clae2	-	clae2 → clae2_desc clae2_desc → clae2
Operador_ orgánico	id	id_departamento	id → {establecimiento, razón_social, id_provincia, id_departamento}
Provincia	id	-	id → provincia provincia → id
Departamento	id	id_provincia	id → {departamento, id_provincia} {departamento, id_provincia} → id
Establecimiento_ productivo	id	clae2 id_departamento	id → {clae2, proporción_mujeres, id_departamento}

Es_producido_por	producto, id_op_or	-	{producto,ip_op_or} → {producto,ip_op_or} La única DF es la trivial por lo tanto su conjunto minimal es esta

Antes de la justificación de por qué cada una está en 3FN se justificará por qué están en 1FN y en 2FN. Están en 1FN ya que el dominio de los atributos incluye solo valores atómicos y no tiene relaciones dentro de relaciones o relaciones como valores de atributos dentro de tuplas. Para el caso de 2FN se analizará relación por relación. En el caso de Producto, Clae, Operador orgánico, Provincia y Establecimiento productivo vemos que las K y CK están conformadas por un único elemento. En el caso de Es_producido_por la única DF es la trivial por lo tanto cumple. En el caso de Departamento, la única DF a analizar es {departamento, id_provincia} → id. En este caso la CK es {departamento, id_provincia}. La dependencia es completa ya que se cuenta con nombres de departamentos repetidos en distintas provincias, por lo que si se elimina id_provincia para dos valores iguales de departamento se tiene id distinto, y porque las provincias tienen más de un departamento, por lo que si se elimina departamento para dos valores de id_provincia iguales se tiene id distinto. Así se ve que la dependencia es completa. Por lo tanto, todas las tablas están en 2FN. Ahora sí, justifiquemos por qué están en 3FN

Están en 3FN ya que están en 2FN y además si se analizan las DFs se puede ver que o todos los lados izquierdos de las DFs son SK o que los lados derechos son atributos primos, cumpliendo así con la definición. En el caso de Es_producido_por al ser la trivial está en 3FN.

Importación de los datos

Una vez finalizado el proceso de limpieza de las tablas provistas se importó la información correspondiente a las relaciones propuestas. Esta importación se realizó mediante consultas SQL. Vale aclarar que si bien las tablas pasaron por un proceso de mejora de la información y de aumento de calidad hubieron ciertas situaciones que tuvimos que manejar a mano.

Para la relación Departamento se importaron de la tabla localidades los atributos id_departamento como id, departamento e id_provincia.

Para Provincia se importaron de la tabla localidades los atributos, id_provincia como id y provincia.

Para Establecimiento productivo se importaron todos los atributos presentes en la tabla establecimiento productivo después de la limpieza.

Para CLAE se importaron todos los atributos de la tabla de clae.

Para Producto se importó el atributo producto de la tabla de operadores orgánicos y se lo ordenó por orden alfabético creciente ya que luego se tuvo que agregar a mano las clae correspondientes a cada producto y este órden facilitaba la tarea.

Para Operadores orgánicos en primer lugar se importaron los atributos id, establecimiento, razón_social, departamento, id_provincia de la tabla operadores orgánicos. Luego, se agregó el id_departamento comparando si los nombre del departamento y la

provincia descritos para cada operadores orgánicos eran nombres de algunos de los departamentos-provincias de la República Argentina. Sin embargo, esta comparación no resultó suficiente ya que como se observó antes, algunos valores del atributo departamentos de operadores orgánicos no son verdaderos nombres de departamentos, sino por ejemplo, de ciudades no cabeceras o pueblos pertenecientes a algún departamento. Para solucionar este problema, a los valores a los cuales no se les había agregado la id_departamento en la primera comparación, se los comparó nuevamente, ahora con nombres de asentamientos en vez de departamentos. Haciendo estas dos comparaciones se llegó a añadir los valores de id_departamento para más de un 90% de la tabla. También se decidió que la parte de la tabla a la cual no se le había podido asignar el valor de id_departamento fuera descartada ya que no permitía la posibilidad de generar una combinación (JOINs) con los datos de la tabla Departamento puesto que se pierde la propiedad de clave foránea del atributo id_departamento.

Para la Relación Produce importamos los atributos id como id_op_or y producto de la tabla de operadores orgánicos.

Análisis de datos

Ejercicios

h) i) El reporte solicitado se encuentra representado en el archivo “H1_producto_provincia_ordenado.csv” en la carpeta “TablasEjercicios”.

h) ii) La Clae2 más frecuente en el Data Frame df_Establecimiento_productivo es 1, cuya descripción es: "Agricultura, ganadería, caza y servicios relacionados".

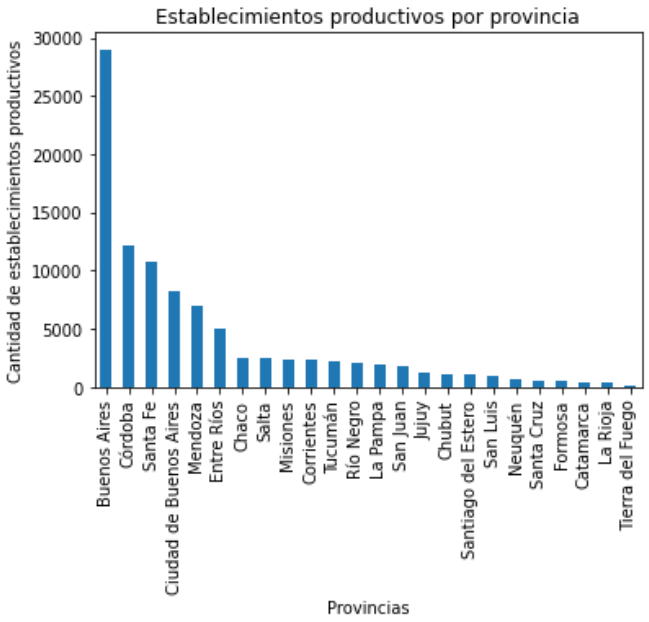
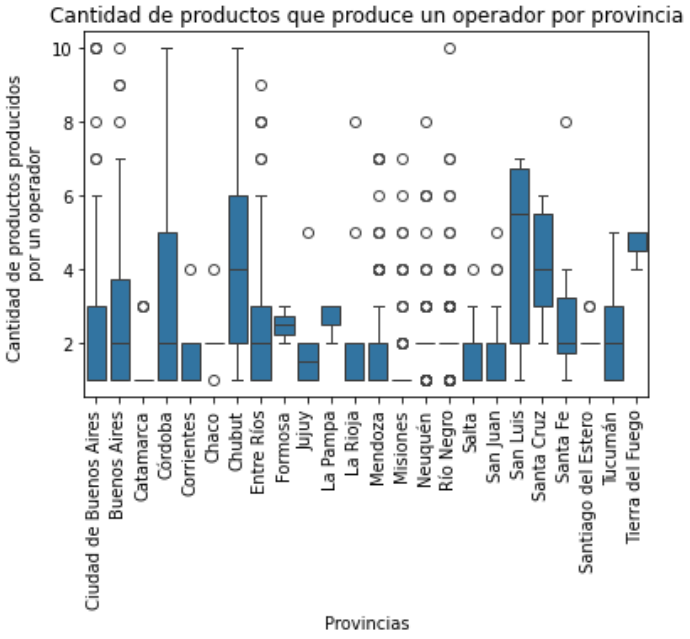
h) iii) El producto más producido es la caña de azúcar y en el archivo “H3_2_provincia_departamento_producto_mas_producido.csv” en la carpeta “TablasEjercicios” se detalla en qué provincia-departamento se producen estas.

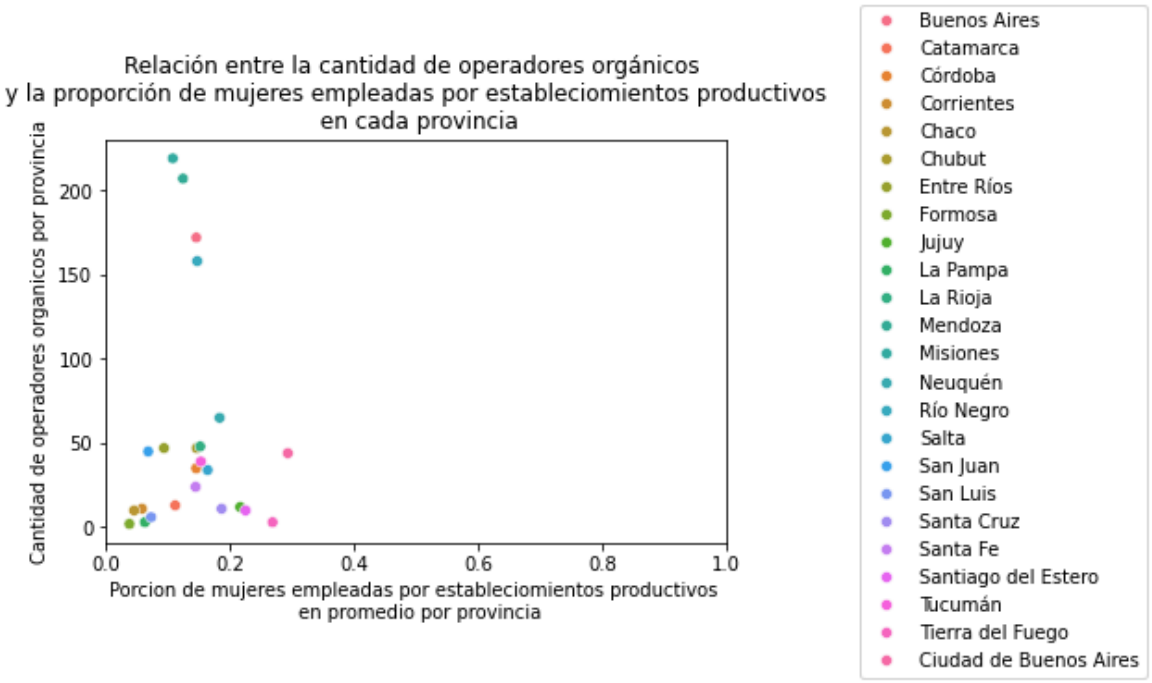
h) iv) Si, existen departamentos que no presentan Operadores Orgánicos Certificados, son un total de 312 departamentos y se encuentran detallados en el archivo “H4_2_departamentos_sin_op_or.csv” dentro de la carpeta “TablasEjercicios”.

h) v) Las respuestas a cada una de las preguntas enunciadas en este ejercicio se encuentran detalladas para cada una de las provincia en el archivo “H5_tabla_prop_mujeres_prom_desvio_mayor_menor.csv” dentro de la carpeta “TablasEjercicios”.

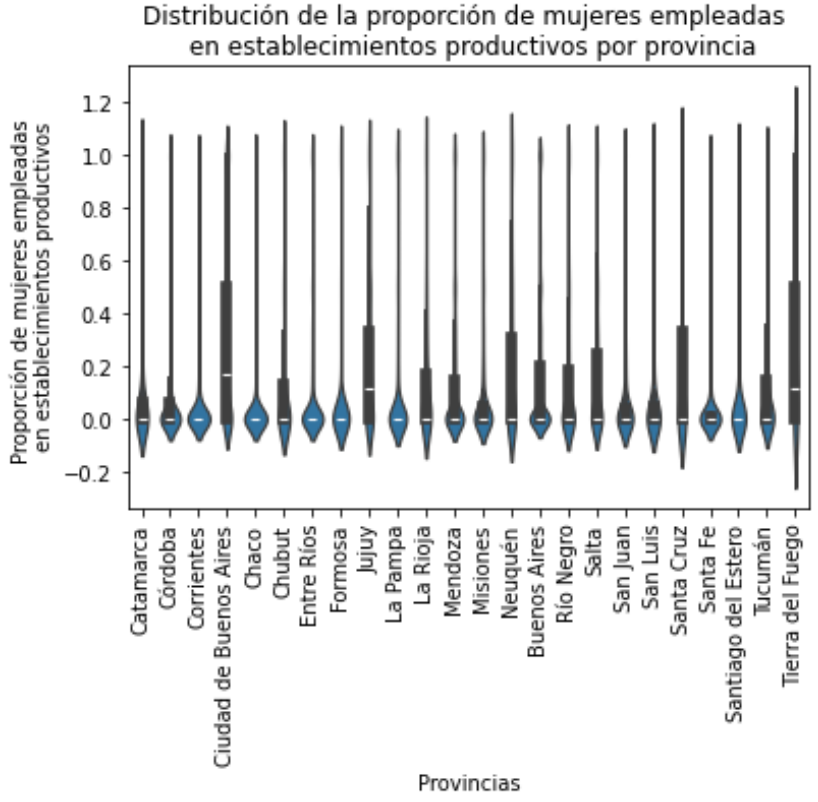
h) vi) La muestra solicitada se encuentra detallada en el archivo “H6_cantidades_depto_prov.csv” dentro de la carpeta “TablasEjercicios”.

Visualización

Cantidad establecimientos productivos por provincia

Cantidad de productos por operador

Para una mayor claridad se decidió colocar el gráfico que limita hasta 10 productos
Cantidad de establecimientos de operadores orgánicos certificados de cada provincia y la proporción de mujeres empleadas en establecimientos productivos de dicha provincia



Distribución de los datos correspondientes a la proporción de mujeres empleadas en establecimientos productivos en Argentina



Conclusiones

Ahora que se cuenta con el análisis de datos se está en condiciones de determinar el objetivo del trabajo: la existencia de cierta relación entre el desarrollo de la actividad orgánica y la proporción de mujeres empleadas en establecimientos productivos de las provincias. No obstante, analizar la existencia de una relación sólo considerando los datos se hizo verdaderamente difícil dada la falta de información contundente. Esto demuestra algo fundamental y es que no se pueden analizar datos sin entender el contexto que acompaña al modelo. Los datos por sí solos no dicen nada, hay que saber cómo interpretarlos de acuerdo al contexto del que provienen. Sin embargo, se pudo extraer una serie de conclusiones.

En primer lugar, se hace evidente la marginación de la mujer en el sector agrícola de la República Argentina ya que vemos cómo en promedio la proporción de mujeres en establecimientos productivos es menor a 0.5. Realmente son pocas aquellas provincias donde la proporción de mujeres es mayor al promedio de la proporción de todo el país, el cual es realmente bajo.

En segundo lugar, también es posible visualizar como en provincias que destacan por su cantidad de operadores orgánicos la proporción es menor a 0.20 (es posible detectar este comportamiento en el gráfico Cantidad de establecimientos de operadores orgánicos certificados de cada provincia y la proporción de mujeres empleadas en establecimientos productivos de dicha provincia). No solo ocurre esto, sino que también si se analiza con detenimiento el gráfico de Cantidad de productos por operador en conjunción con el mencionado previamente se puede visualizar como en provincias donde los productores orgánicos producen 2 o más productos la proporción de mujeres es también menor a 0.20.

Esta situación es sorprendente ya que si se considera la cantidad de puestos de trabajo que genera el sector agrícola dada a su importancia a nivel económico del país la única explicación posible de este fenómeno está ligada a la perpetuación de estereotipos machistas que descalifican a las mujeres por su condición de ser mujer. Incluso se ve como grandes productores o provincias con una amplia cantidad de estos fallan en la integración de la mujer a su matriz laboral. Integración que dada la cantidad de puestos a cubrir debería ser sencilla y cuyo incumplimiento solo pone en evidencia la existencia de un sesgo contra las mujeres.

Este suceso es una clara demostración de la falta de educación en materia de género que se presenta a lo largo de todo el territorio argentino y lo poco que se hace en materia política para luchar contra esta inequidad. Incluso, si vemos en detalle las fuentes, la provincia que mayor proporción de mujeres presenta es la Ciudad Autónoma de Buenos Aires lo cual no sorprende dado que es el sector del país que más en contacto está con cuestiones de género. Y también desilusiona porque muestra la falta de interés de los políticos en asegurar la inclusión de la mujer más allá de CABA.

Así, se puede concluir que no hay una relación ya que hay una marginación absoluta de la mujer en el desarrollo de la actividad orgánica del país lo que no permite visualizar qué es lo que ocurriría a mayor proporción.