

Evaluaciones Agropecuarias de los municipios de Cundinamarca y Boyacá

Andrea Nataly Hernández Fuentes, Edgar Felipe Torres Ortegón ,
Helmy Andrea Moreno Navarro, y Samuel Mantilla Velásquez

Universidad Sergio Arboleda

Talento Tech

Objetivo

Objetivo General

Realizar un análisis integral de la biodiversidad, conservación ambiental y los ecosistemas en las regiones de Cundinamarca y Boyacá, utilizando técnicas avanzadas de análisis de datos y modelos predictivos. A través de la exploración, análisis y modelado de datos, buscamos identificar patrones significativos, áreas de interés prioritarias y proponer recomendaciones fundamentadas para la gestión sostenible y la conservación eficaz de la biodiversidad en estas regiones.

Objetivos Específicos

Esta fase tiene como objetivo realizar una exploración inicial de los datos disponibles relacionados con la biodiversidad, conservación ambiental y los ecosistemas en las regiones de Cundinamarca y Boyacá. La adquisición de datos de calidad es fundamental para un análisis efectivo.

Índice

Objetivo General	1
Objetivos Específicos	1
Identificación de fuente de datos	3
Documentación Limpieza de Datos	4
Descripción de datos originales	4
Registro de Cambios	5
Rutas	6
Diccionario de datos	7
Exploración inicial de los datos	8
Descripción general del dataSet	8
Datos Categóricos	8
Datos Numéricos	9
Exploración inicial	9
Datos Categóricos	9
Datos Numéricos	15

Identificación de fuente de datos

En la exploración de la fuente de datos se tomo como referencia la información contenida en la pagina <https://agronet.gov.co> de datos libre de la entidad publica **AgroNet**, la cual cuenta con herramientas estadísticas de sectores como: Agrícola, Pecuario, Precios Comercio, Créditos, insumos entre otros, ver 1 Para el caso de estudio se decidió trabajar

Figura 1

Herramientas estadísticas AgroNet



con datos Agrícolas correspondientes a Reportes de *Evaluaciones Agropecuarias - EVA* y *Anuario Estadístico del Sector Agropecuario*, la base de dato seleccionada fue el reporte más actualizado a la fecha '07-2023' como indica en la figura

La información consultada se discriminó mediante el filtro de departamento (*Boyacá y Cundinamarca*), y teniendo en cuenta los temas de interés *Biodiversidad, Conservación Ambiental y Ecosistemas*, en criterio de cantidad de datos y/o registros ligado a la fecha de publicación procurando que los datos no excedan los cinco años de actualización.

Figura 2

Relación de la base de datos seleccionada del sector agrícola

Reporte: Evaluaciones Agropecuarias - EVA y Anuario Estadístico del Sector Agropecuario	
Base Agrícola EVA de 2019 a 2022 - Fecha de publicación 07072023	Descargar
Informe preliminar Análisis de Resultados EVA primer semestre de 2022	Descargar
Anuario digital estadístico del año 2019 - 2021	Descargar
Base Agrícola EVA de 2019 a 2021 - Fecha de publicación 22042022	Descargar

Documentación Limpieza de Datos

Descripción de datos originales

Las evaluaciones agropecuarias municipales (EVA) son la base de información y conocimiento sobre la oferta productiva agropecuaria de los municipios del país, una operación estratégica para la generación de estadísticas y la toma de decisiones en el sector, en articulación con el Sistema Nacional Unificado de Información Rural Agropecuaria (SNUIRA). En este sentido, se diseñan módulos temáticos que se desarrollan de manera secuencial, en un ciclo de diez años, partiendo de un módulo básico que se recolecta anualmente, así como módulos rotativos recogidos periódicamente a lo largo del ciclo.

La base de datos original contenían información de los 32 departamentos de Colombia, en una primera aproximación a lo datos, basándonos en los objetivos del proyecto se decide filtra la información de los departamentos de **Boyacá y Cundinamarca**, dicha información contiene 18337 registros la cual contenía información de : [Código Dane departamento, Departamento, Código Dane municipio, Municipio, Desagregación cultivo, Cultivo Ciclo del cultivo, Grupo cultivo, Subgrupo, Año Periodo, Área sembrada (ha), Área cosechada (ha), Producción (t), Rendimiento (t/ha), Código del cultivo, Nombre científico del cultivo, Estado físico del cultivo], esta se suministraba en un archivo (.xlsx), el cual se relaciona a

continuación  20230704_BaseEVA_Agrícola20192022

Registro de Cambios

Basado en una exploración inicial de los datos hemos adoptado la convención de *Pascal Case* para nombrar nuestros campos y variables. Además, hemos realizado cambios en los nombres para que sean más descriptivos y fáciles de entender, este cambio se puede visualizar en el diccionario de datos ver tabla 1.

Por otro lado en la revisión de valores nulos/vacíos se encontró que el campo **Cultivos** poseía 66 valores nulos, al hacer una revisión de los campos que se relacionaban con este se encontró que este campo se relacionaba con la columna **desagregacionCultivo** allí se visualiza que es una categoría general **Otros** en esos registros, se decide cambiar este campo vacío de cultivo por el registro del campo de desagregacionCultivo.

Entre los cambios significativos que hemos implementado, destaca la sustitución de los registros en la columna ‘periodo’. Dado el contexto de los datos, esta columna se refiere a un periodo semestral que cumple con las condiciones ambientales adecuadas para el cultivo de ciertos alimentos. Con el objetivo de evitar redundancias en la información, hemos realizado la siguiente modificación

Se realiza el cambio de los registros ‘**Periodo**’ por su valor alfabético correspondiente al semestre del año

- A = Primer semestre
- B = Segundo semestre
- C = año completo

Figura 3

Limpieza de datos columna Periodo

df[['Anio', 'Periodo']].head(10)		
✓	0.0s	
Anio	Periodo	
0	2019	2019
1	2019	2019A
2	2019	2019A
3	2019	2019A
4	2019	2019A
5	2019	2019A
6	2019	2019A
7	2019	2019A
8	2019	2019A
9	2019	2019B

(a) *Columna Periodo sin modificaciones*



df[['Anio', 'Periodo']].head(10)		
✓	0.0s	
Anio	Periodo	
0	2019	C
1	2019	A
2	2019	A
3	2019	A
4	2019	A
5	2019	A
6	2019	A
7	2019	A
8	2019	A
9	2019	B

(b) *Columna Periodo modificada*

Una vez finalizada la primera etapa de limpieza de los datos de Evaluaciones Agropecuarias, creamos un conjunto de datos denominado ***EvaluacionAgro.csv*** en formato CSV delimitado por ‘;’. Este conjunto de datos está relacionado en la documentación de la tabla 1.

Con el objetivo de gestionar la documentación, hemos decidido crear un repositorio para el control de versiones de los datos y la documentación del proyecto.

Rutas

-  Evaluaciones-Agropecuarias-Boyacá-Cundinamarca
-  Bases de Datos

Diccionario de datos

Cuadro 1

Diccionario de datos de Evaluaciones agropecuarias de los departamentos de Boyacá y Cundinamarca, Fecha de actualización: 07-07-2023

Titulo	EVALUACIONES AGROPECUARIAS MUNICIPALES - 2019 A 2022		
Url	Link	Boyacá y Cundinamarca	
Nombre del dataset	EvaluacionAgro.csv	Separador : ','	
Filas,Columnas	(18337, 18)	Cantidad de registros: 18337	
Documentación			
Atributo	Muestra	Tipo	Observación
codDaneDpto	15	Int	Muestra el código del Departamento
Dpto	Boyacá	String	Muestra el nombre del Departamento
codDaneMunicipio	15022	Int	Muestra el código del Municipio
Municipio	Almeida	String	Muestra el nombre del Municipio
desagregacionCultivo	Tomate Invernadero	String	Se refiere a la subdivisión o separación de los datos agrícolas en categorías más específicas o detalladas.
Cultivo	Tomate	String	Muestra el nombre del cultivo
cicloDelCu1tivo	Transitorio	String	Indica el periodo de tiempo desde la siembra hasta la cosecha de un cultivo. Cuando se clasifica un cultivo como permanente o transitorio hace referencia a la duración del cultivo y a su relación con el suelo y la planta.
grupoCultivo	Hortalizas	String	Muestra categorías más amplias o generales de cultivos que comparten características comunes.
SubGrupo	Hortalizas de Futo	String	Muestra categorías más específicas o detalladas dentro de un grupo de cultivos más amplio.
Anio	2019	Int	Muestra el año del cultivo
periodo	2019A	String	Indica el periodo del cultivo el cual se encuentran años que finalizan con la letra A o la letra B, donde los años que finalizan con la letra A se refieren al primer semestre y los años que finalizan con la letra B se refieren al segundo semestre.
areaSembradaHa	2	Float	Representa la extensión de tierra (en hectáreas) en la que se ha sembrado un cultivo específico durante un período de tiempo determinado.
areaCosechadaHa	2	Float	Representa la extensión de tierra (en hectáreas) que ha sido efectivamente cosechada para recolectar el cultivo.
produccionTon	136	Float	Indica la cantidad total de producto agrícola (en toneladas) que se ha obtenido de la cosecha.
rendimientoTonHa	68	Float	Representa la medida de la eficiencia de la producción agrícola y se expresa como la cantidad de producto agrícola (en toneladas) obtenida por unidad de área de tierra sembrada (hectáreas). Es decir, es la producción por hectárea de tierra sembrada.
codCultivo	1052902	Int	Muestra el código del cultivo
nombreCientificoCultivo	Lycopersicon esculentum	String	Muestra el nombre científico del cultivo
estadoFisicoCultivo	En fresco	String	Describe el estado físico del cultivo en el momento de la medición o recolección.

Exploración inicial de los datos

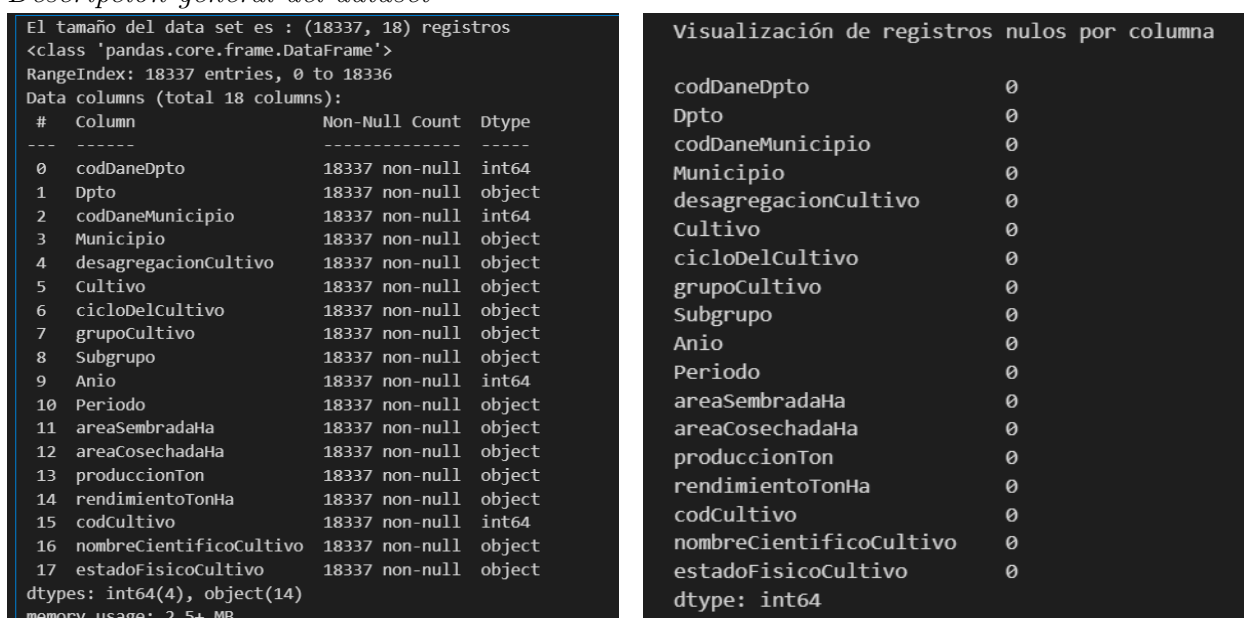
A continuación se presenta la exploración inicial del conjunto de datos acerca del dataset **EvaluacionAgro.cvs** correspondiente a evaluaciones agropecuarias en los departamentos de Boyacá y Cundinamarca

Descripción general del dataSet

El dataset cuenta con 18337 registros y 18 columnas entre ellas contienen valores de tipo categóricos y numéricos

Figura 4

Descripción general del dataset



(a) Relación de columnas y tipo de dato

(b) valores nulos

En una primera visualización de los datos se llegó relacionar el tipo de dato por columna de la siguiente manera:

Datos Categóricos

- codDaneDpto
- Dpto
- codDaneMunicipio
- Municipio
- desagregacionCultivo
- Cultivo
- cicloDelCultivo
- grupoCultivo
- Subgrupo
- nombreCientificoCultivo
- estadoFisicoCultivo
- codCultivo
- Periodo

Datos Numéricos

- areaSembradaHa
- produccionTon
- Anio
- areaCosechadaHa
- rendimientoTonHa

Exploración inicial

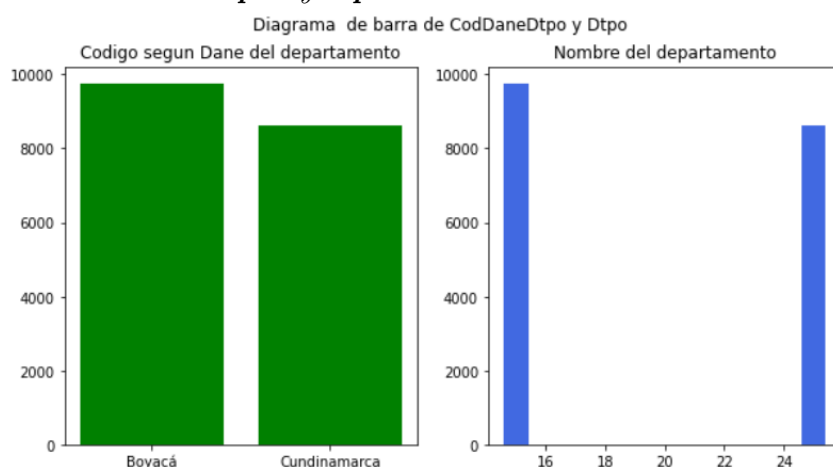
Debido a que tenemos diferentes tipos de datos, realizaremos una exploración inicial de cada uno de los campos. Con base en la información encontrada, tomaremos acciones para mejorar el conjunto de datos.

Datos Categóricos

Para los datos categóricos de opto por realizar diagramas de barras a continuación se presenta cada de uno de ellos.

Figura 5

Gráfico de barras de `codDaneDpto` y `Dpto`



Dado que las gráficas de frecuencia representadas anteriormente muestran que este campo corresponde al mismo registro, donde una columna representa el nombre y la otra el código asociado a dicho nombre, es pertinente evitar la redundancia en los datos. Para mantener la calidad, eficiencia y claridad de los datos, se recomienda quedarnos únicamente con el campo ‘Dpto’, ya que este es más descriptivo para nuestro dataset.

De manera similar, observamos un comportamiento similar con las variables ‘Municipio’ y ‘codMunicipio’. Siguiendo el mismo criterio de selección, consideramos que la variable ‘Municipio’ es más descriptiva para nuestros datos.

Al hacer un conteo de los datos únicos en esta columna ‘Municipio’, notamos que habían un gran número de datos únicos por lo tanto se tomo la decisión de establecer un filtro según el departamento y de este modo observar los municipios en un diagrama de barras.

Figura 6

Municipios de Boyacá

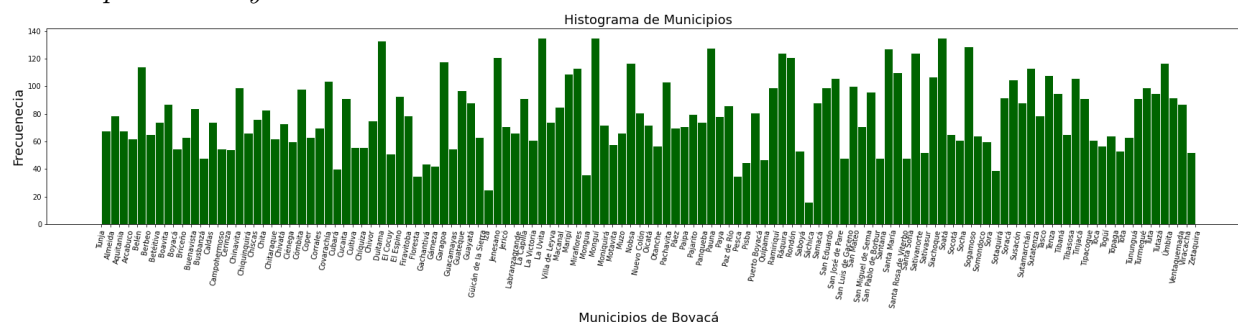
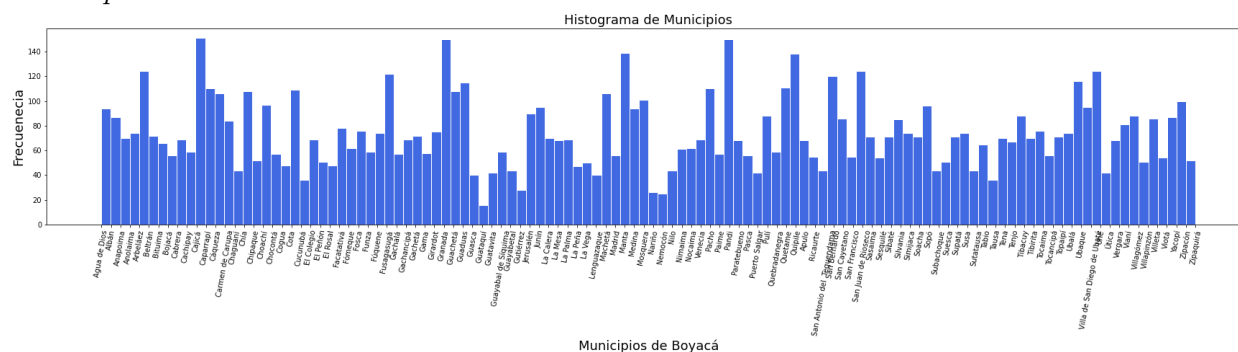


Figura 7

Municipios de Cundinamarca



Dado que las columnas ‘desagregacionCultivo’ y ‘codCultivo’ corresponden al nombre y la otra el código asociado a dicho nombre, es pertinente evitar la redundancia en los datos. Para mantener la calidad, eficiencia y claridad de los datos, se recomienda quedarnos únicamente con el campo ‘desagregacionCultivo’, ya que este es más descriptivo para nuestro dataset

Figura 8

desagregacionCultivo y codCultivo

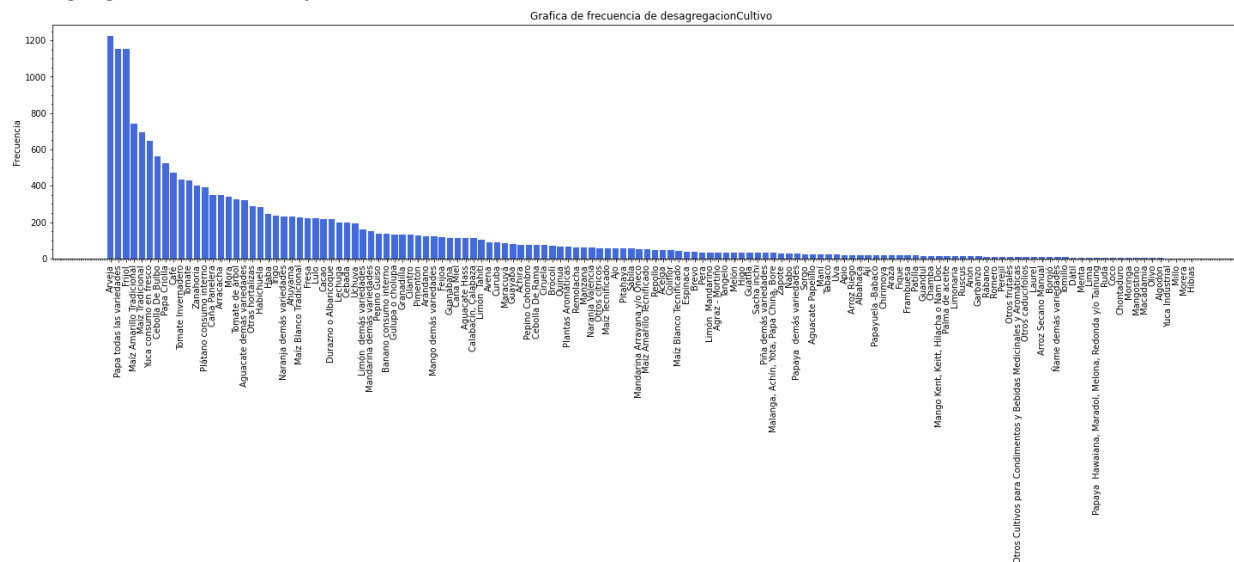


Figura 9

Gráfica de los cultivos de los departamentos de Boyacá y Cundinamarca

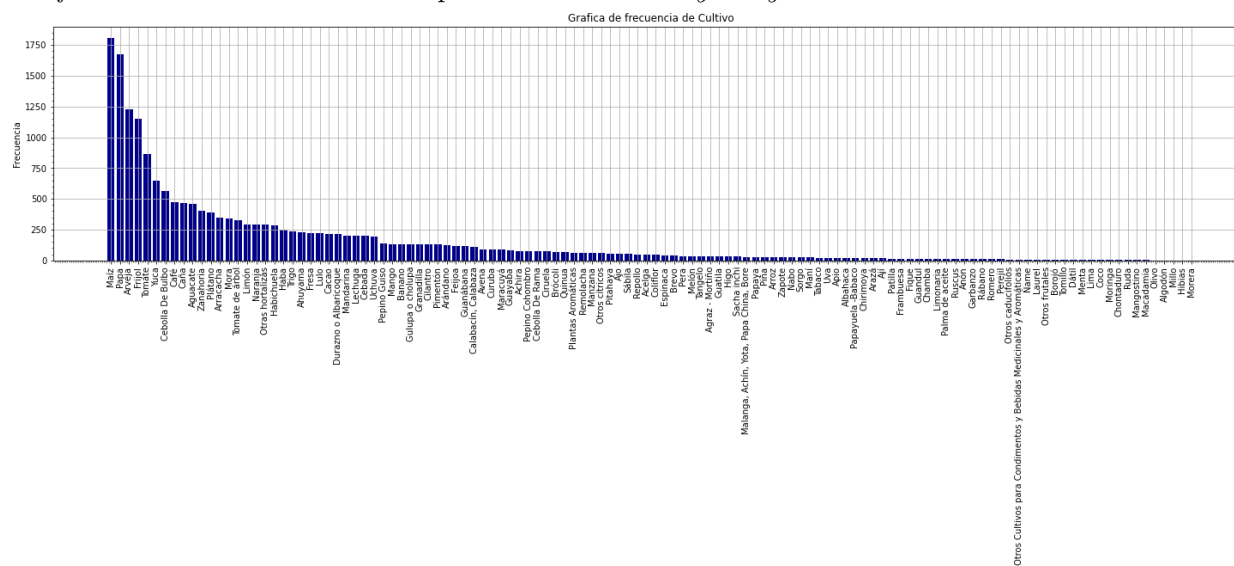
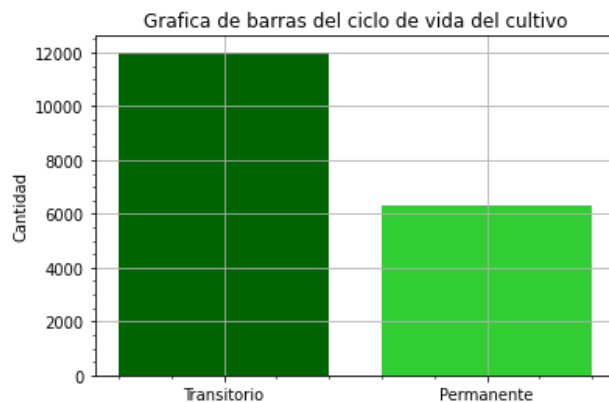


Figura 10

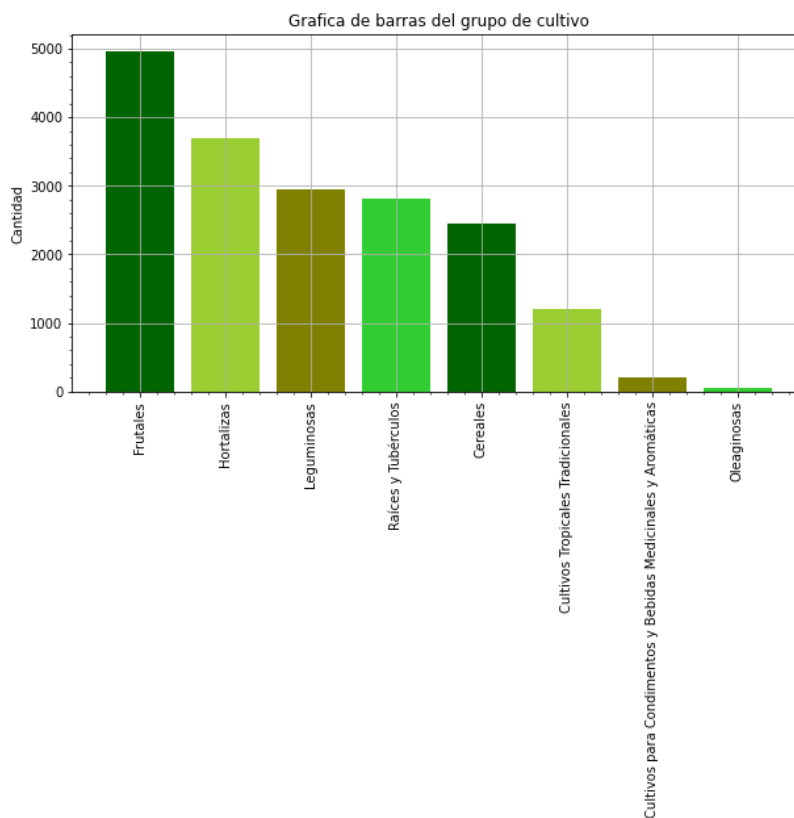
Diagrama de barras de ciclo de vida del cultivo cicloCultivo



Note. Cuando se clasifica un cultivo como permanente o transitorio hace referencia a la duración del cultivo y a su relación con el suelo y la planta.

Figura 11

Diagrama de barras de grupo de cultivos 'grupoCultivo'



Note. Muestra categorías más amplias o generales de cultivos que comparten características comunes

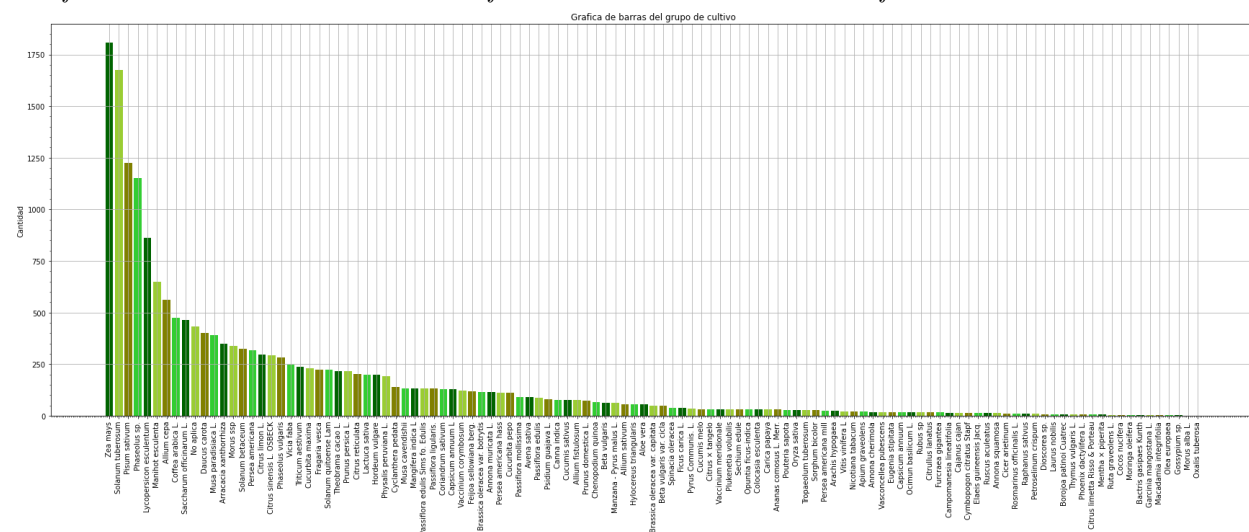
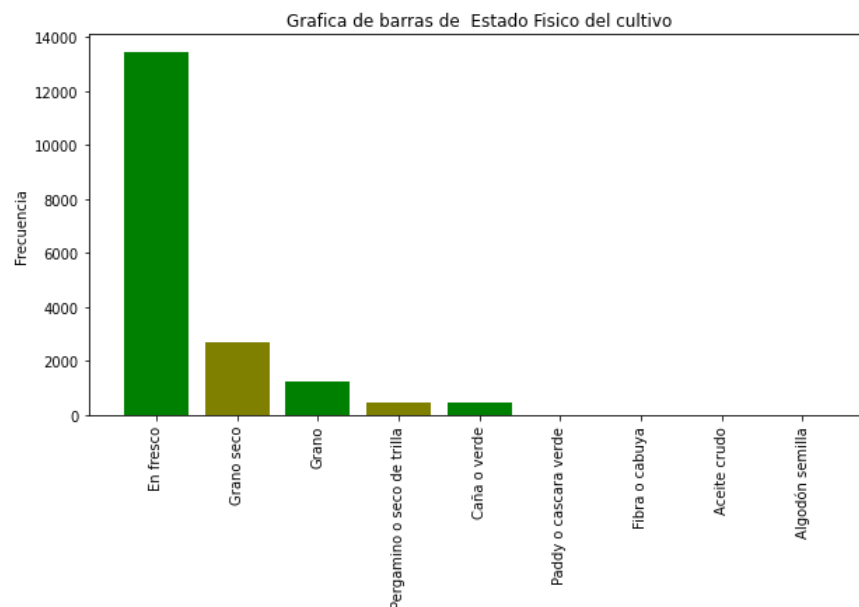


Figura 14

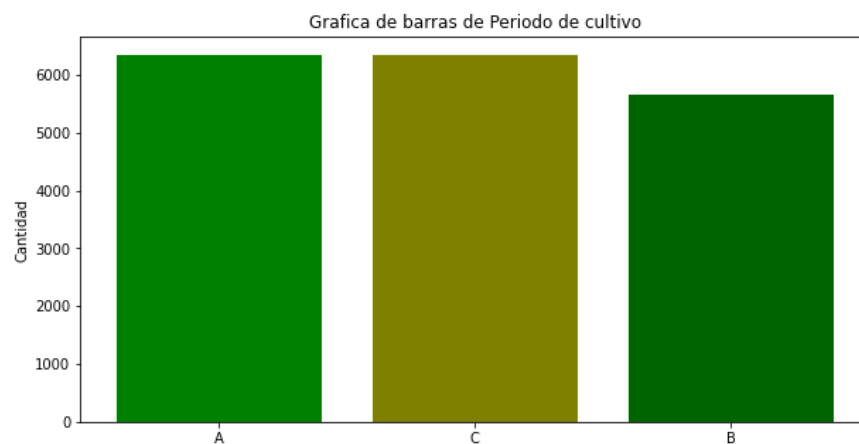
Diagrama de barras del estado físico del cultivo estadoFisicoCultivo



Note. Describe el estado físico del cultivo en el momento de la medición o recolección.

Figura 15

Gráfico de barras del 'Periodo'



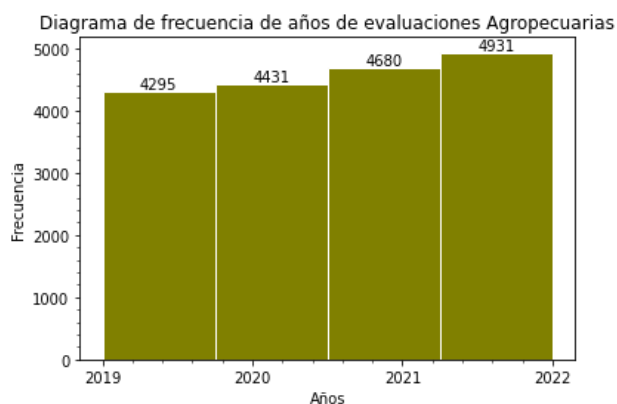
Note. Corresponde al periodo del semestre siendo A el primero periodo, B el segundo y C correspondiente a los dos periodos

Datos Numéricos

A continuación se presenta los diagrama de frecuencia para los datos numéricos

Figura 16

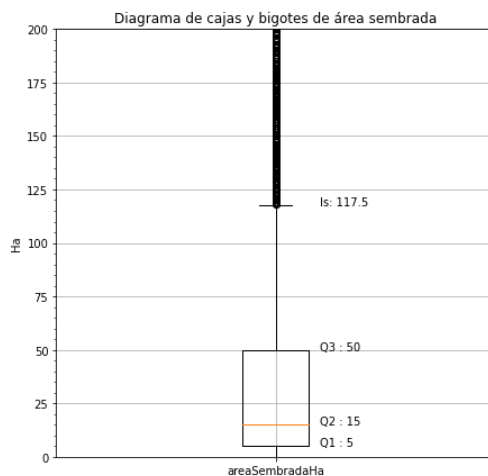
Histograma año del cultivo



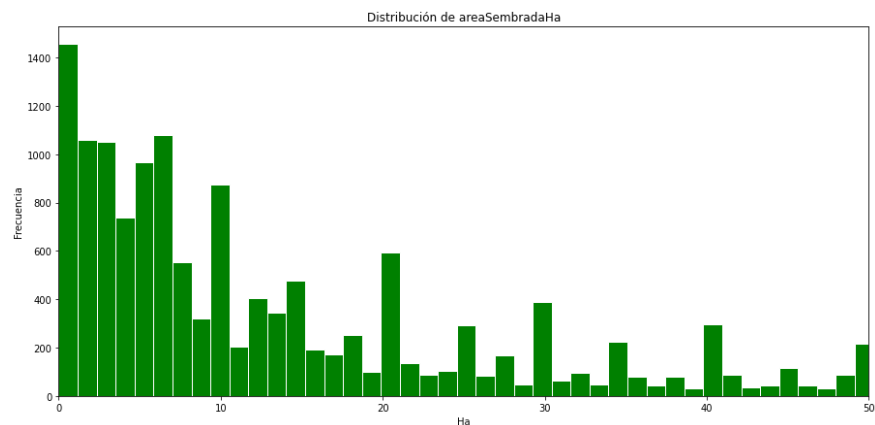
Para analizar los datos, comenzamos graficando un diagrama de caja (boxplot) con el objetivo de identificar valores atípicos y determinar el valor del bigote superior. Esto nos permitió acotar los datos y visualizar su distribución. En todos los casos, se observó una distribución sesgada hacia la izquierda.

Figura 17

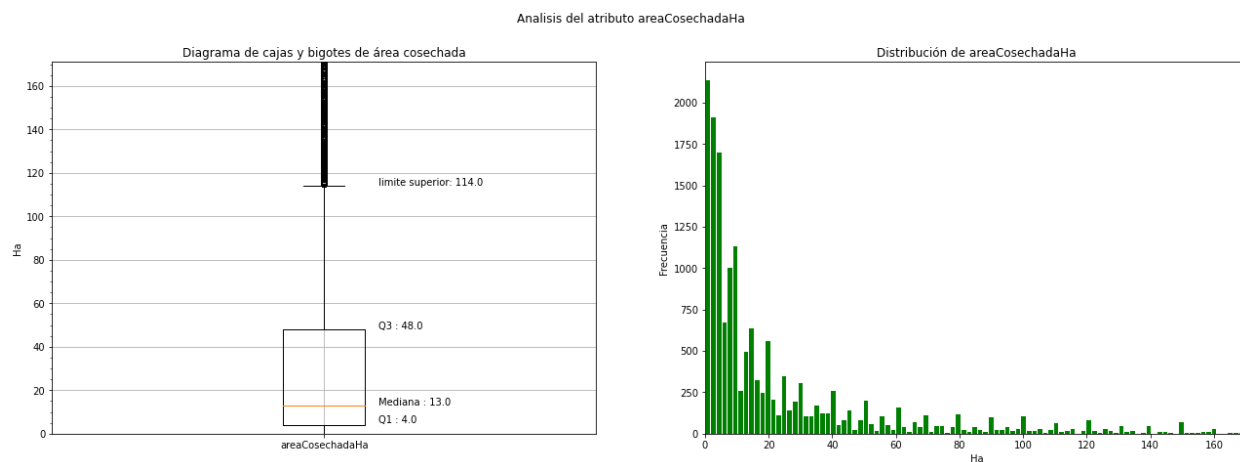
Boxplot 'areaSembradaHa'



Note. Se determino la distribución y las tendencias centrales, mediante esto se indican los valores en la gráfica

Figura 18*Distribución de 'areaSembradaHa'*

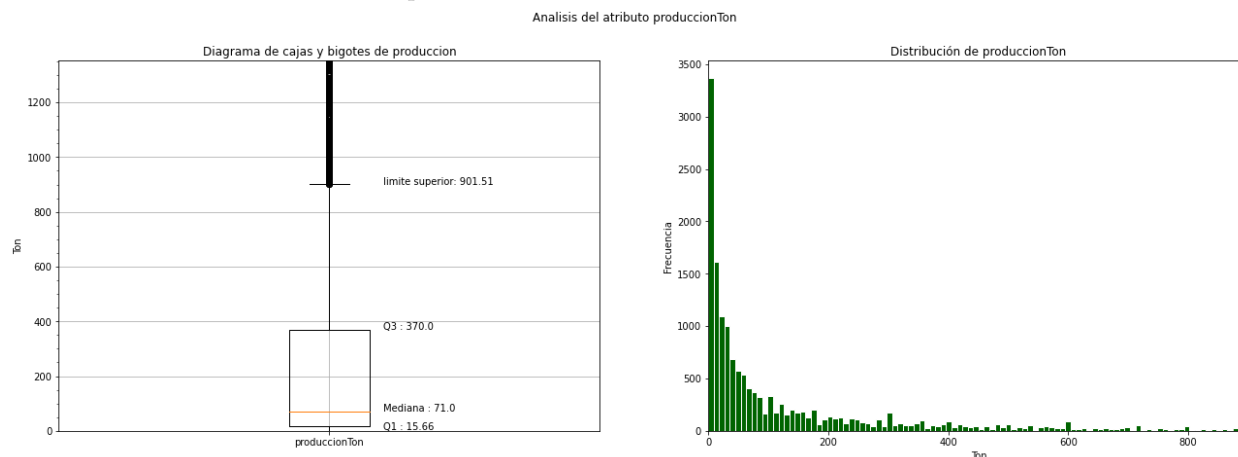
Note. Representa la extensión de tierra (en hectáreas) en la que se ha sembrado un cultivo específico durante un período de tiempo determinado.

Figura 19*Distribución de los datos de 'areaCosechadaHa'*

Note. Representa la extensión de tierra (en hectáreas) que ha sido efectivamente cosechada para recolectar el cultivo.

Figura 20

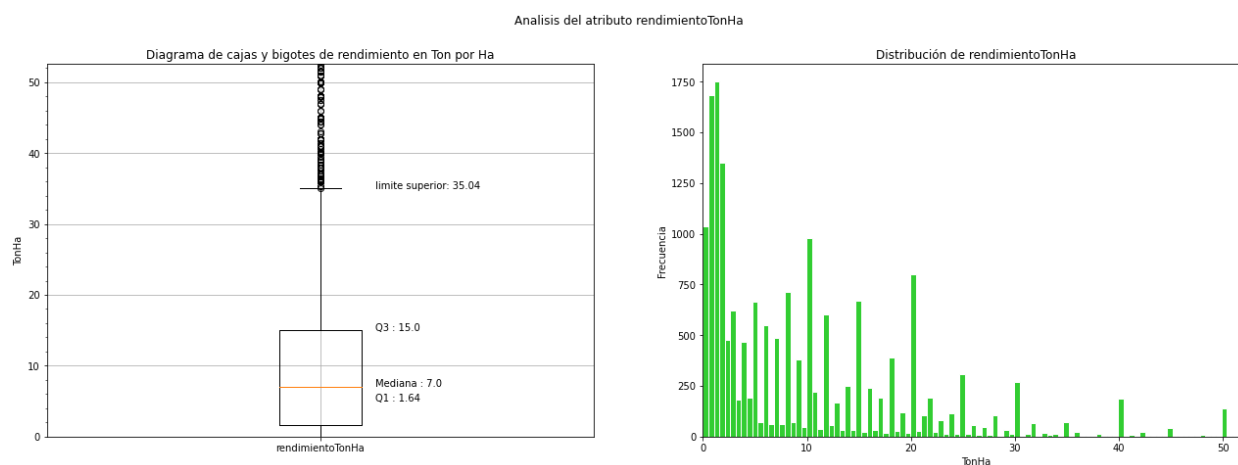
Distribución de los datos de 'produccionTon'



Note. Indica la cantidad total de producto agrícola (en toneladas) que se ha obtenido de la cosecha.

Figura 21

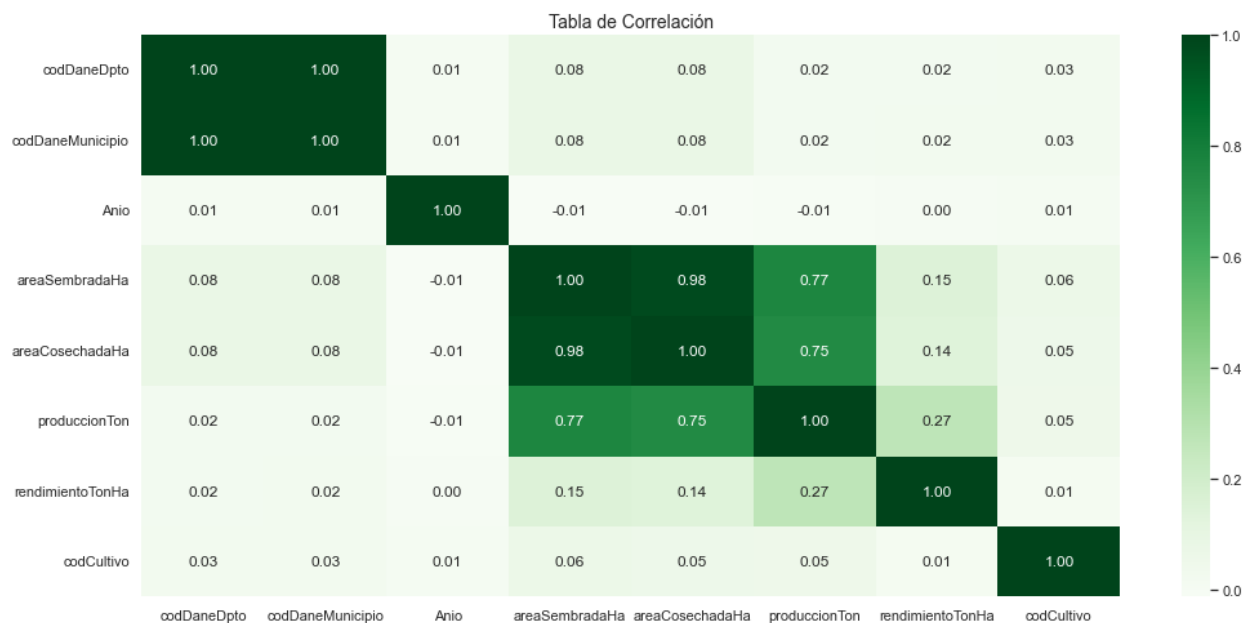
Distribución de los datos de 'rendimientoTonHa'



Note. Representa la medida de la eficiencia de la producción agrícola y se expresa como la cantidad de producto agrícola (en toneladas) obtenida por unidad de área de tierra sembrada (hectáreas). Es decir, es la producción por hectárea de tierra sembrada.

Figura 22

Tabla de correlación de variables numéricas



Note. La tabla de correlación se evidencia que las variables más tentativas para ser usadas como los **features** son: produccionTon, areaSembradaHa y areaCosechadas, ya que se evidencia una correlación más alta.