PROYECTO FINAL DE BIGDATA Y MACHINE LEARNING

ANÁLISIS DE PRODUCCIÓN DE GRANOS EN ARGENTINA, PRINCIPALMENTE TRIGO Y SOJA.

- REALIZADO POR: IVAN GONZALO TAPIA
- INSTITUCIÓN: UNIVERSIDAD PROVINCIAL DEL SUDESTE
- MATERIA: BIG DATA Y APRENDIZAJE DE MÁQUINA
- DOCENTE: ING. VALENTÍN BARCO
- LUGAR: PUNTA ALTA
- FECHA: 20 DE DICIEMBRE DE 2023

- REPOSITORIO PROYECTO: GITHUB PROYECTO FINAL BIGDATA Y MACHINE LEARNING
- REPORTE: INFORME EN WORD
- NOTEBOOK: COLAB

JUSTIFICACIÓN DEL INFORME Y PROPÓSITOS

NOS CENTRAREMOS EN VARIOS CONJUNTOS DE DATOS QUE INCLUYEN INFORMACIÓN DE DIFERENTES CULTIVOS COMO TRIGO, SOJA, AVENA, MAIZ, CENTENO, CENTRANDONOS PRINCIPALMENTE EN LA PRODUCCIÓN DEL TRIGO Y LA SOJA QUE SON LOS PRINCIPALES PROTAGONISTAS EN ARGENTINA.

1- INTRODUCCIÓN
2- METODOLOGÍAS, LIBRERIAS UTILIZADAS, CARGA DE DATOS, FUNCIÓNES PARA SANITIZAR LOS DATOS Y PARA ORGANIZAR DATA SETS DESORGANIZADOS.
3- ANÁLISIS EXPLORATORIO DE DIFERENTES GRANOS.
4- ANÁLISIS EXPLORATORIO PROFUNDO DE TRIGO Y SOJA.
5- CREACIÓN DE MODELOS KMEANS DE SKLEARN.
6- REFLEXIÓN FINAL.

1. Introducción

El proyecto tiene como objetivo principal realizar un análisis exhaustivo del rendimiento y producción de dos de los cultivos más significativos en Argentina: trigo y soja. La iniciativa trata de proporcionar una comprensión detallada de los patrones de producción a nivel nacional.

Objetivos del proyecto

Producción agrícola Argentina.

Análisis de Producción: Evaluar y comparar el rendimiento anual de diferentes granos, poniendo mayor énfasis y haciendo un análisis más exhaustivo para el trigo y la soja analizando su nivel de producción en todas las provincias y departamentos de Argentina.

Exploración Geoespacial: Utilizar algoritmos de clustering para identificar diferentes niveles de producción de trigo y soja en las diferentes regiones geográficas de Argentina.

Visualización Interactiva: Generar gráficos y visualizaciones interactivas que permitan a los usuarios explorar los datos de producción y rendimiento de manera intuitiva y amigable.

Justificación

Argentina es un actor clave en la producción agrícola a nivel mundial, y el trigo y la soja son fundamentales para nuestra economía. Entender las variaciones en la producción a nivel regional es esencial para interiorizarse en los niveles de producción de cada región en nuestro país.

Alcance del Proyecto

El alcance abarca la totalidad del territorio argentino, examinando datos a nivel de provincias y localidades. Se emplearán técnicas de análisis de datos, visualización y algoritmos de Clustering para ofrecer una visión integral y detallada de los patrones de producción de trigo y soja.

2. Metodologías, librerías utilizadas, carga de datos, funciones para sanear los datos y para organizar Data sets desorganizados.

En esta sección, se realizaron los pasos iniciales para el desarrollo del proyecto, que incluyen la importación de librerías, carga y preprocesamiento de datos.

Librerías necesarias para el desarrollo del proyecto

Ver figura 1.

```
import os
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import plotly.graph_objects as go
import plotly.express as px
from plotly.subplots import make_subplots

from sklearn.cluster import KMeans, MeanShift, DBSCAN

from sklearn.preprocessing import LabelEncoder
from sklearn.neighbors import NearestNeighbors
from sklearn.metrics import silhouette_score
from sklearn.cluster import estimate_bandwidth
```

Figura 1: Librerías de Python importadas para el desarrollo del proyecto.

Carga de datos

Se cargaron los datos desde varios archivos .csv y .xlsx, los cuales fueron extraídos de fuentes oficiales del gobierno, todos los Data sets están publicados en la página oficial del gobierno.

Funciones para sanear datos y para organizar data sets

Se creo la función ‘sanitizar_dataframe(df)’ que sanea el Data set eliminando filas con valores NaN y convierte las columnas numéricas al tipo correspondiente. Y la función ‘organizar_datos(df)’ que está diseñada para procesar un data set que proviene de un archivo Excel con un formato específico. El archivo tiene una columna combinada que contiene múltiples campos separados por comas, y el objetivo de la función es dividir esta columna en

varias columnas individuales, asignar nombres apropiados a cada columna y luego combinarlas con el data set original, formando un nuevo data set con el formato esperado y apto para ser analizado.

3. Análisis exploratorio de diferentes granos.

En esta sección se analiza la producción de diferentes granos a lo largo del tiempo en Argentina aproximadamente desde 1923 hasta 2020, teniendo como referencia la superficie plantada por hectárea, superficie cosechada por hectárea, producción por tonelada y el rendimiento en kilogramo por hectárea.

3.1 Avena

En este primer análisis de avena, un importante cultivo para Argentina, podemos observar cómo fluctúa su siembra, cosecha y producción a lo largo del tiempo, ver figura 2.

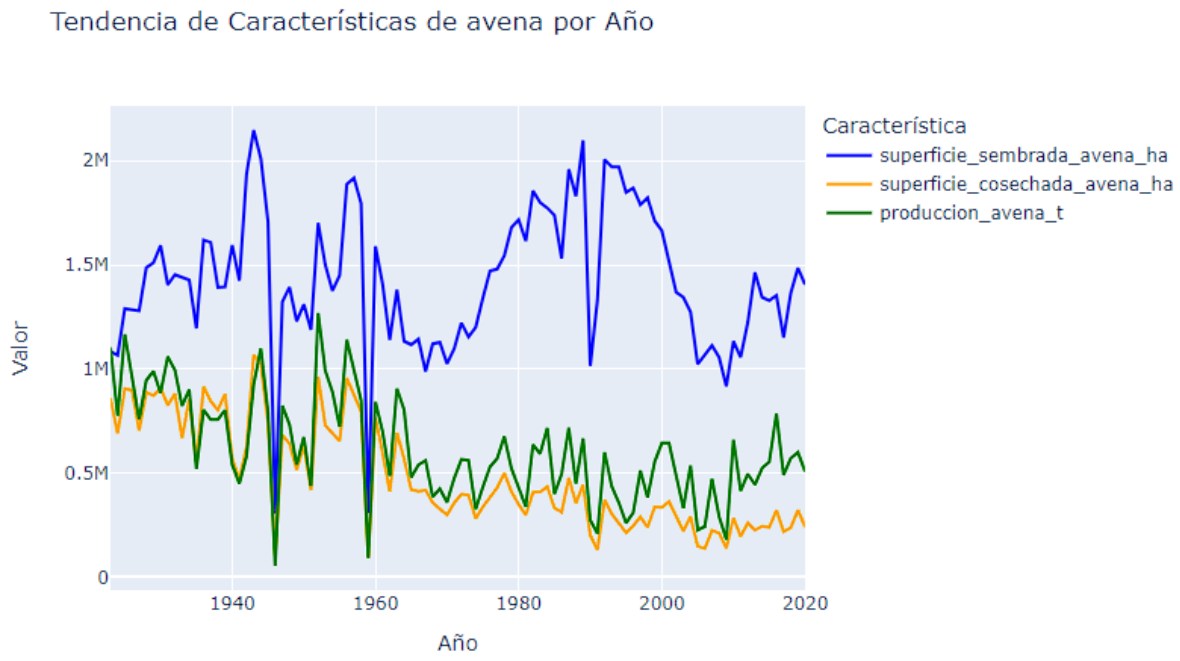

Figura 2: Producción de avena en Argentina desde 1923 hasta 2020.

Podemos observar cómo se comportó la producción de avena en estos años, y dos picos bajistas en los años 1946 y 1959, donde la superficie sembrada para este cultivo no supero los 0.4 millones de hectáreas, estos factores coinciden con fenómenos climáticos y políticos argentinos ya que, en 1946, una sequía afectó gran parte del país, lo que provocó una caída en la producción de cereales, incluidos la avena. Esta sequía fue particularmente grave y afectó gran parte del país lo que provocó una disminución en la producción en todas las provincias productoras de avena (Libro "Clima y sociedad en la Argentina" de Carlos A. Paruelo y otros). Por otro lado, en 1959, una gran inundación afectó a las provincias de Buenos Aires, Entre Ríos, Santa Fe, Córdoba, La Pampa y San Luis principales zonas productoras de avena en Argentina. La inundación de 1959 fue una de las más graves que ha sufrido Argentina en el

siglo XX (https://www.mundoagrario.unlp.edu.ar/article/view/v08n16a11/969). Además de los fenómenos climáticos adversos, en ambos años, Argentina experimentó cambios en sus políticas. En 1946, el gobierno de Juan Domingo Perón implementó una serie de políticas proteccionistas que favorecieron a los cultivos de granos básicos, como el trigo y el maíz. Esto provocó una disminución en la producción de cultivos menos rentables, como es la avena, centeno, etc. En 1959, el gobierno de Arturo Frondizi implementó una serie de reformas económicas que liberalizaron el mercado de granos. Esto provocó una disminución en los precios de los granos, lo que pudo llevar a una reducción en la producción de avena. (https://www.argentina.gob.ar/interior/archivo-general-de-la-nacion).

A pesar de los desbalances y picos bajistas en la superficie sembrada de avena en Argentina, el rendimiento del cultivo a lo largo del tiempo siempre se mantuvo tendencia alcista desde 1959. En 1923, la producción de avena por hectárea era de 1.200 kilogramos., y en 2018, el rendimiento aumentó a 2.400 kilogramos por hectárea, un aumento de rendimiento del 100% ver figura 3, teniendo en cuenta que la superficie dedicada a este cultivo fue decreciendo desde 1959 hasta lo que podemos visualizar del 2020.

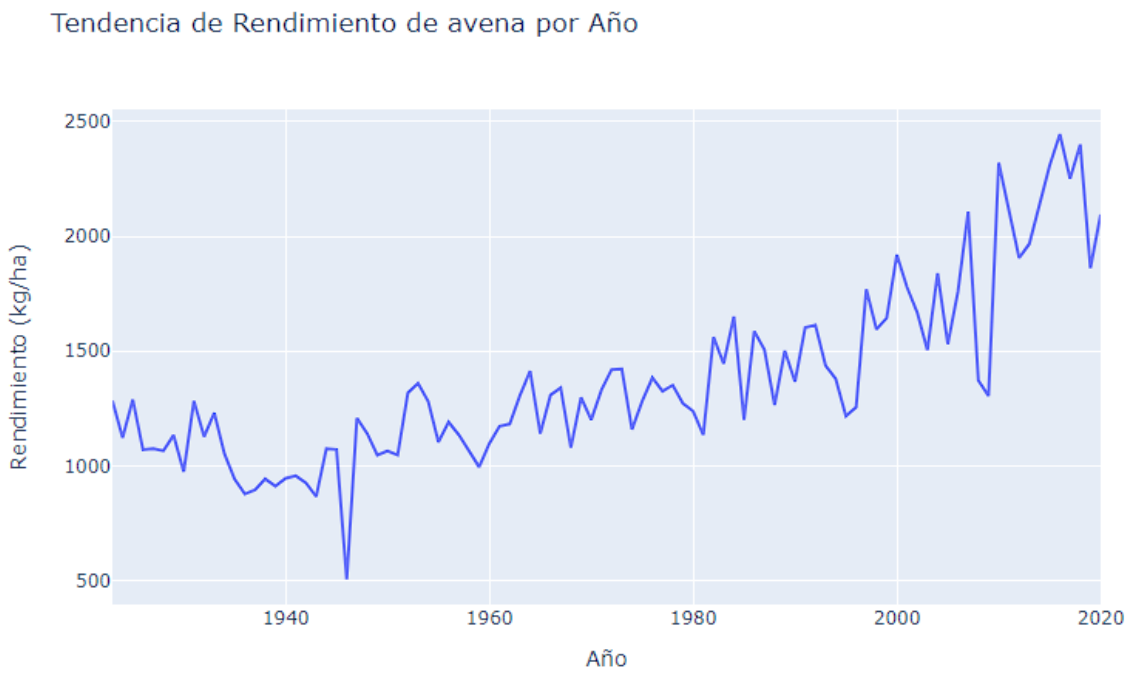

Figura 3: Rendimiento de kilogramos avena cosechados por hectárea desde 1923.

3.2 Maíz

En el análisis de producción de maíz en Argentina observamos una tendencia general de crecimiento en la superficie sembrada y cosechada de maíz, con un único episodio considerable de disminución en 1949. En 1949, Argentina sufrió una sequía severa que afectó a gran parte del país. La sequía de 1949 fue particularmente grave en la región pampeana, la principal zona productora de maíz en Argentina. Además, provocó una disminución de la producción de maíz en Argentina en un 70%. En 1949, la producción de maíz en Argentina fue de solo 850.000 toneladas, la cifra más baja registrada en la historia del país.

En las últimas dos décadas, la producción de maíz en Argentina ha experimentado un crecimiento notable. En 1923, la producción de maíz en Argentina fue de 5 millones de toneladas. En 2000, la producción de maíz en Argentina fue de 15 millones de toneladas. En 2020, la producción de maíz en Argentina fue de 59 millones de toneladas, un aumento del 392% en relación con 1923, ver figura 4.

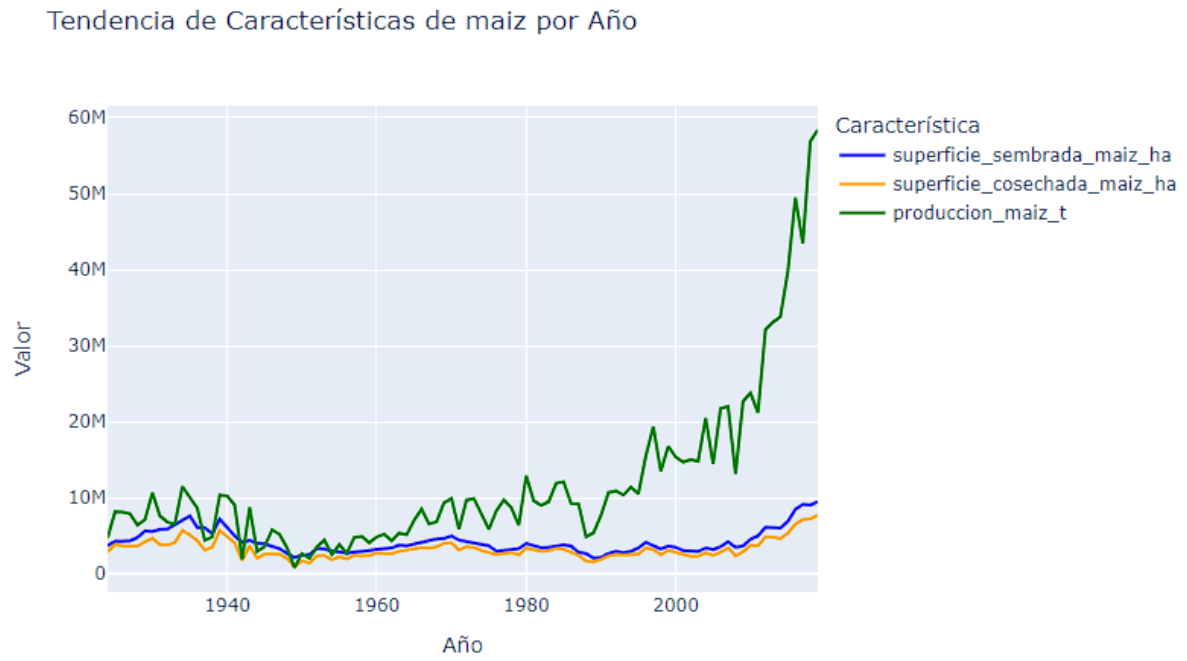

Figura 4: Producción de maíz en Argentina desde 1923 hasta 2020.

El gráfico de rendimiento del maíz en Argentina muestra una tendencia general de aumento, con un único episodio de disminución en 1949. Entre 1923 y 1970, el rendimiento del maíz se mantuvo estable oscilando entre 1.000 y 2.450 kg por hectárea. El rendimiento más bajo se registró en 1949, con 890 kg por hectárea.

A partir de 1970, el rendimiento del maíz comenzó a aumentar gradualmente. En 1980, el rendimiento alcanzó los 3.900 kg por hectárea. En 1990, el rendimiento aumentó a 4.500 kg por hectárea. En 2000, el rendimiento alcanzó los 6.000 kg por hectárea. En 2020, el rendimiento alcanzó los 7.500 kg por hectárea, esto representa un aumento del 90% en el rendimiento en 50 años. Ver figura 5.

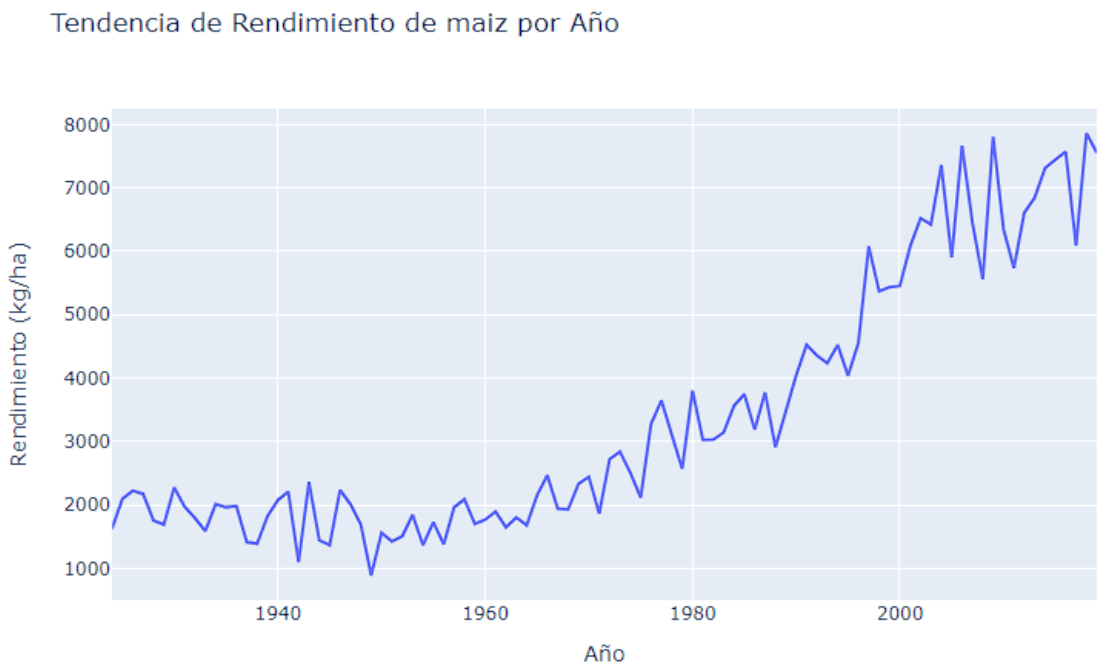

Figura 5: Rendimiento de maíz desde 1923 hasta 2020(kg. Cosechado por hectárea)

3.3 Girasol

El análisis de producción de girasol en Argentina muestra una tendencia general de aumento los primeros años, con un pico en 1998 y un descenso posterior hasta 2020. Entre 1969 y 1975, la producción de girasol en Argentina rondaba el millón de toneladas por año, lo que representa un 9% de la producción mundial. A partir de 1975, la producción comenzó a aumentar gradualmente, alcanzando los 2 millones de toneladas en 1985, lo que representa un 15% de la producción mundial. En 1998, alcanzó un nivel récord de 7 millones, ver figura 6.

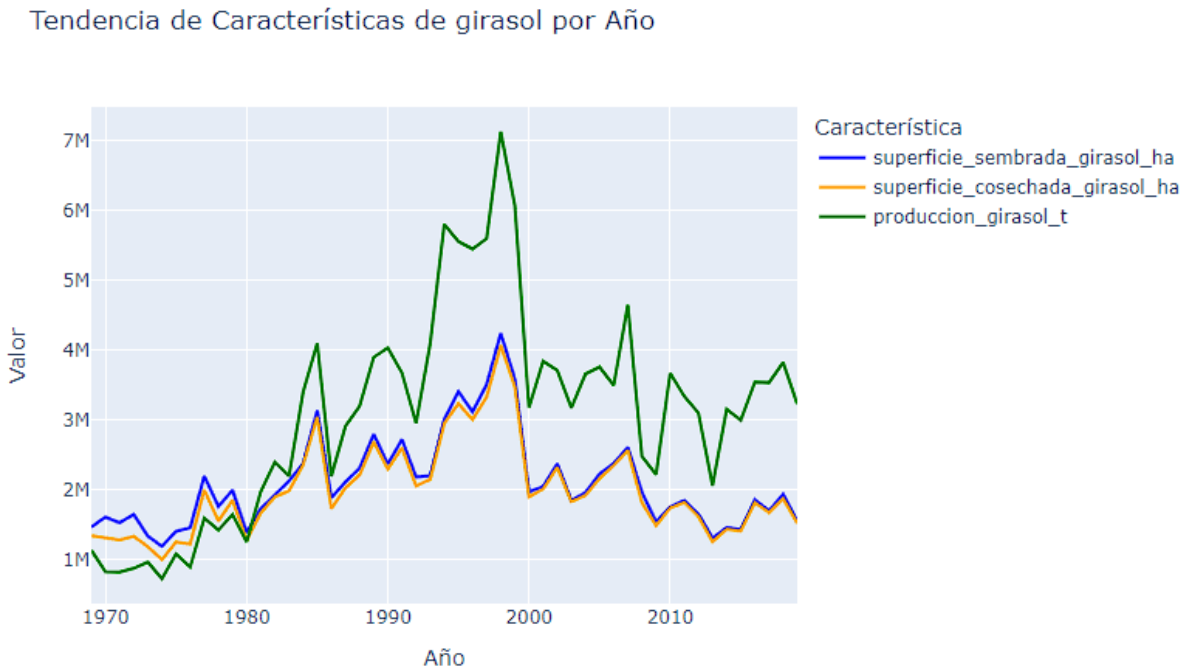

Figura 6: Producción, cosecha y siembra de girasol desde 1969 hasta 2020.

A partir de 1998, la producción de girasol comenzó a disminuir gradualmente. Este descenso se puede atribuir a una serie de factores, entre los que se incluyen la disminución de la demanda de aceite de girasol en el mundo. En la década de 2000, la demanda mundial de aceite de girasol se estabilizó, debido al crecimiento de la producción de aceite de soja en Brasil y Estados Unidos. Esto redujo las oportunidades para los productores argentinos de girasol. Y también los cambios en las políticas agrícolas, ya que, en la década del 2000, el gobierno argentino implementó una serie de políticas agrícolas que favorecieron la producción de otros cultivos, como la soja, el maíz y el trigo.

3.4 Trigo

El análisis de producción de trigo en Argentina muestra una tendencia general de aumento, con un pico en 2019 y una leve disminución posterior en 2020, ver figura 7.

Entre 1923 y 1970, la producción de trigo en Argentina osciló entre los 3 y 10 millones de toneladas por año. En este período, la producción representó un promedio del 12% de la producción mundial. A partir de 1980, la producción de trigo comenzó a aumentar significativamente. En 2019, la producción alcanzó un nivel récord de 20 millones de toneladas, lo que representó un 16% de la producción mundial.

El aumento de la producción de trigo en Argentina se puede atribuir a una serie de factores, entre los que se incluyen, el aumento de la demanda mundial de trigo. La demanda mundial de trigo ha aumentado significativamente en las últimas décadas, debido al crecimiento de la población y al aumento de los ingresos. En 2019, la demanda mundial de trigo representó un 24% del consumo mundial de cereales. Los precios del trigo han aumentado significativamente en las últimas décadas, lo que ha hecho que la producción de trigo sea más rentable. En 2019, el precio promedio del trigo fue de 250 dólares por tonelada

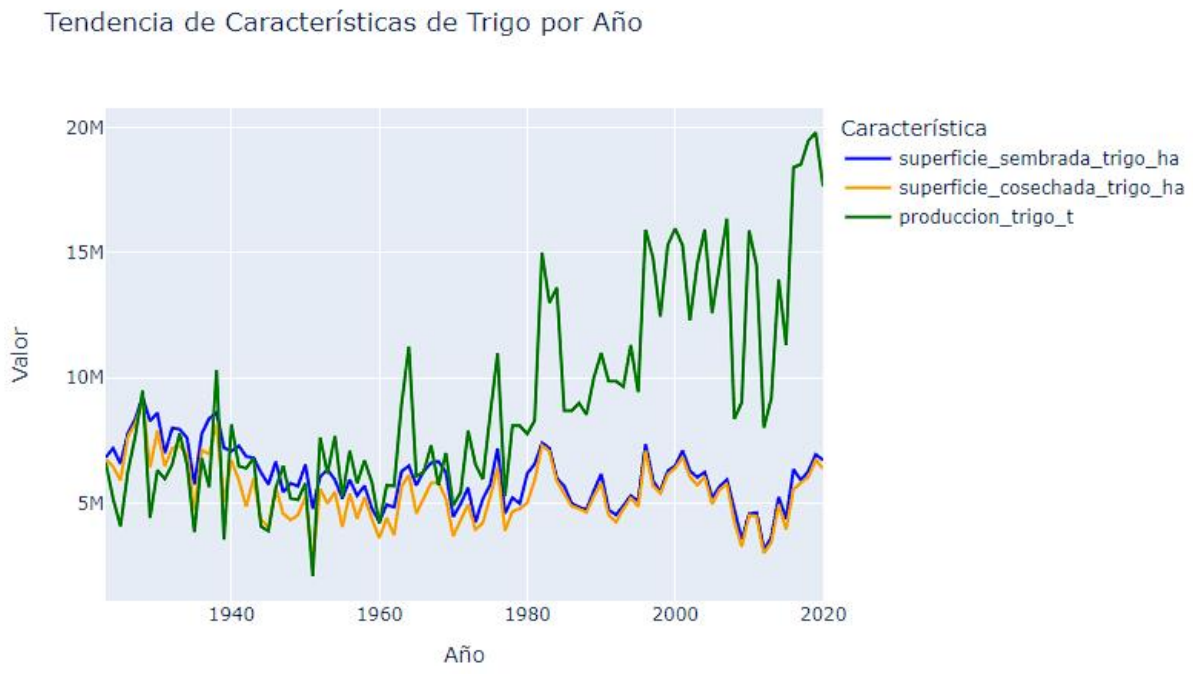

Figura 7: Producción, cosecha y siembra de trigo desde 1969 hasta 2020.

El gráfico de rendimiento de trigo en Argentina muestra una tendencia general de aumento, con un pico máximo en 2010. En 1923, el rendimiento de trigo en Argentina era de 1,000 a 500 kilogramos por hectárea. Este nivel de rendimiento se mantuvo hasta 1940. A partir de 1940, el rendimiento comenzó a aumentar gradualmente. En 1951, el rendimiento alcanzó un pico de 1,450 kilogramos por hectárea, lo que representa un 45% de aumento con respecto al rendimiento de 1923. Sin embargo, en 1939 y 1951, se observaron dos picos muy bajos, con rendimientos de menos de 500 kilogramos por hectárea, ver figura 8.

A pesar de estos desbalances, la tendencia de rendimiento siempre fue alcista. En 1980, el rendimiento se situó en torno a 1,500 kilogramos por hectárea, lo que representa un 50% de aumento con respecto al rendimiento de 1951. En 2000, el rendimiento aumentó a 2,500 kilogramos por hectárea, lo que representa un 75% de aumento con respecto al rendimiento de 1980. Entre 2010 y 2020, el rendimiento alcanzó un pico máximo de 3,500 kilogramos por hectárea, lo que representa un 140% de aumento con respecto al rendimiento de 2000 y un 350% de aumento con respecto al rendimiento de 1923.

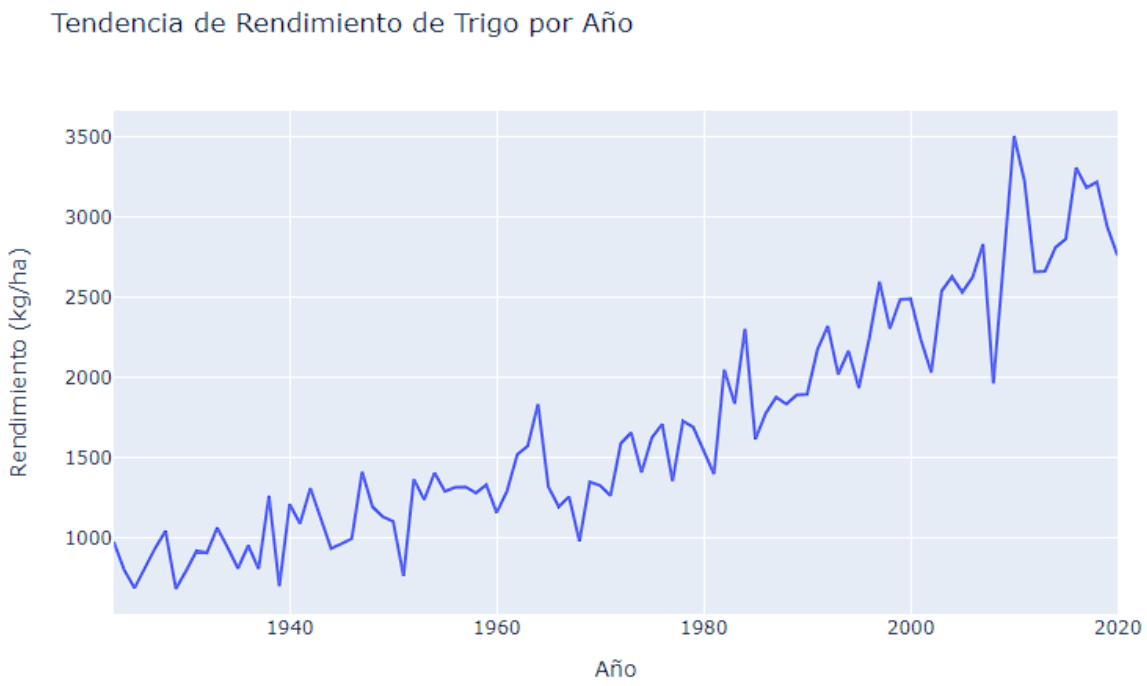

Figura 8: Rendimiento de maíz desde 1923 hasta 2020(kg. Cosechado por hectárea).

3.5 Soja

El gráfico de producción de soja en Argentina muestra una tendencia general de aumento, con un crecimiento exponencial en los últimos años. En 1969, la producción de soja en Argentina era de 36,300 toneladas por año. Este nivel de producción bajísimo se mantuvo hasta 1975. A partir de 1975, la producción comenzó a aumentar exponencialmente. En 1990, la producción alcanzó los 10 millones de toneladas, lo que representa un 275% de aumento con respecto a la producción de 1975. De 1990 a 2000, la producción aumentó a 27 millones de toneladas, lo que representa un 170% de aumento con respecto a la producción de 1990. Y del 2000 a 2020,

la producción aumentó a 50 millones de toneladas, lo que representa un 85% de aumento con respecto a la producción de 2000 y un 500% con respecto a 1990. Un dato importante: en 2020, Argentina se situó como el tercer productor mundial de soja, con una producción de 50 millones de toneladas anuales, ver figura 9.

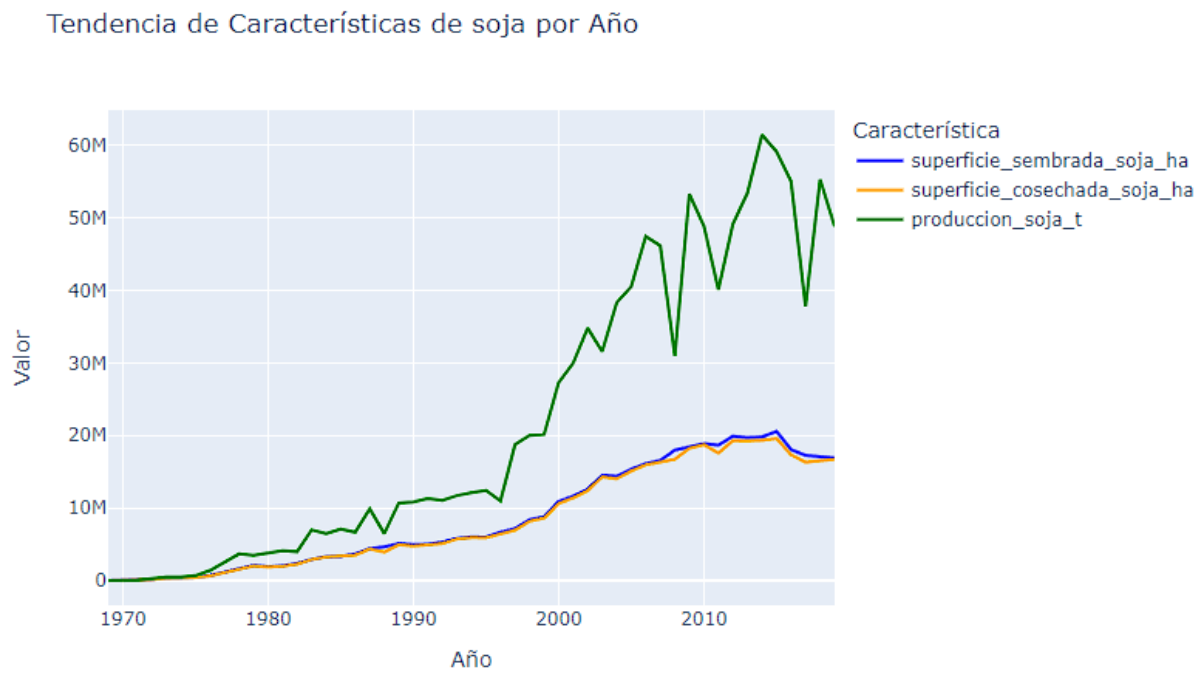

Figura 9: Producción, cosecha y siembra de soja desde 1969 hasta 2020.

En 1970, el rendimiento de soja en Argentina era de entre 1000 y 1500 kilogramos por hectárea. Este nivel de rendimiento se mantuvo hasta 1975. A partir de 1975, el rendimiento comenzó a aumentar gradualmente. En 1990, el rendimiento alcanzó los 2000 kilogramos por hectárea, lo que representa un 50% de aumento con respecto al rendimiento de 1975. De 1990 a 2000, el rendimiento aumentó a 2,500 kilogramos por hectárea, lo que representa un 75% de aumento con respecto al rendimiento de 1990. Y del 2000 a 2020, el rendimiento aumentó a 3,000 kilogramos por hectárea, lo que representa un 20% de aumento con respecto al rendimiento de 2000. El rendimiento de soja en Argentina es superior al promedio mundial, que se sitúa en torno a los 2,500 kilogramos por hectárea. Ver figura 10.

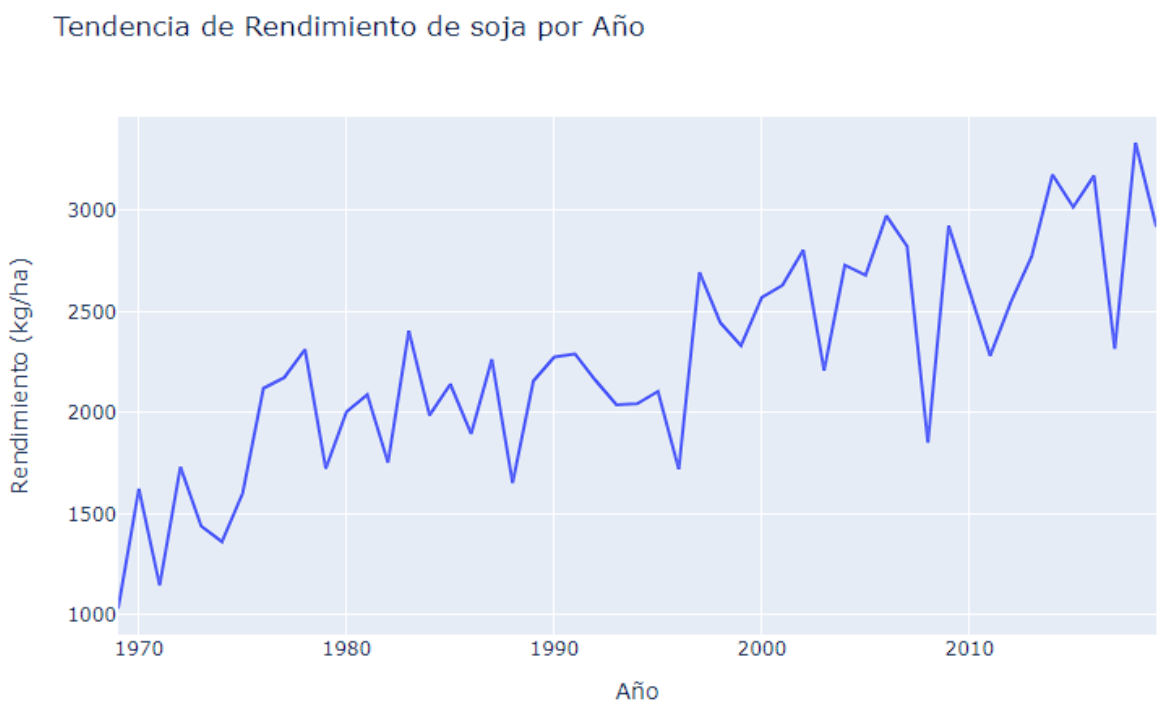

Figura 10: Rendimiento de soja desde 1923 hasta 2020(kg. Cosechado por hectárea).

3.6 Centeno

El análisis de producción de centeno en Argentina muestra una tendencia general de disminución, con un pico máximo de producción en la década de 1950/1960 y un estancamiento en las últimas décadas. En 1923, la producción de centeno en Argentina era de 50,000 toneladas por año. Este nivel de producción se mantuvo hasta 1940. A partir de 1940, la producción comenzó a aumentar gradualmente. En 1950, la producción alcanzó las 500,000 toneladas, lo que representa un 900% de aumento con respecto a la producción de 1940. En 1951, la producción se redujo a 81,000 toneladas, lo que representa una 82% de disminución con respecto a la producción de 1950, en esta disminución de 1951, Argentina sufrió una sequía severa, que afectó a la producción de cereales en todo el país, este patrón lo pudimos observar de igual manera en el trigo.

En la década de 1950, la producción de centeno alcanzó su punto máximo, con una producción media de 750,000 toneladas por año. En la década de 1960, la producción continuó aumentando, con un pico de 1,333,000 toneladas en 1952. A partir de 1970, la producción de centeno comenzó a disminuir gradualmente. Entre 2010 y 2020, la producción se situó en torno a 180,000 toneladas. Ver figura 11.

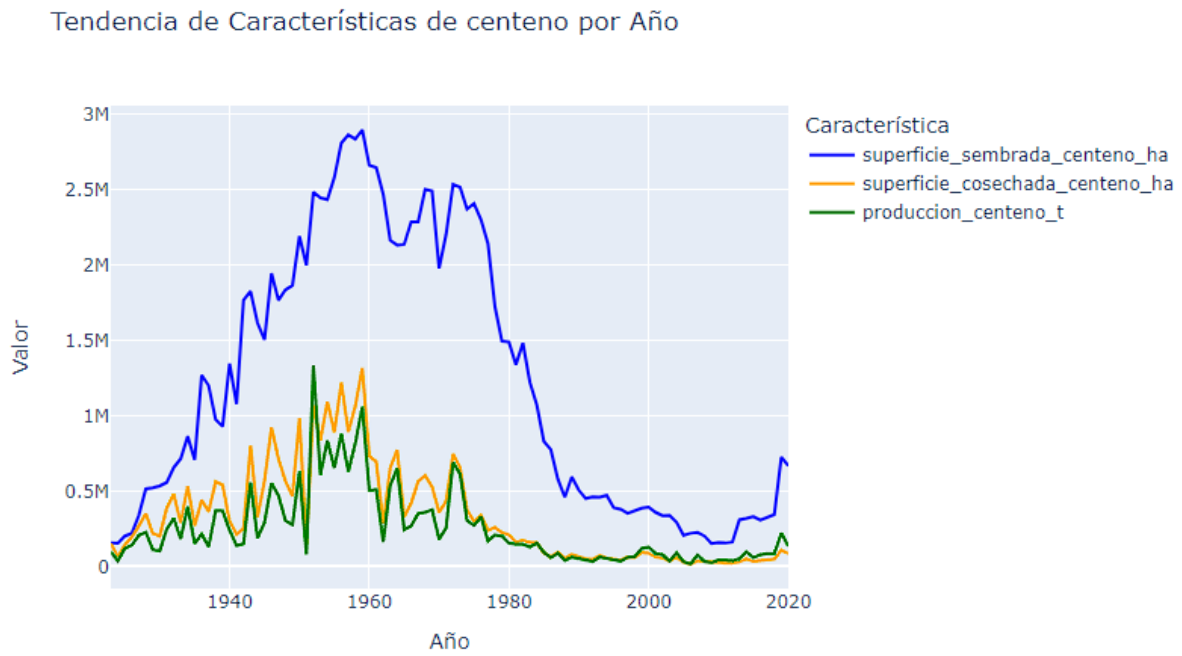

Figura 11: Producción, cosecha y siembra de centeno desde 1969 hasta 2020.

En gráfico de rendimiento de centeno (figura 12) en Argentina muestra una tendencia general de aumento, con un aumento significativo en las últimas décadas. En 1924, el rendimiento de centeno en Argentina era de 600 kilogramos por hectárea. Este nivel de rendimiento se mantuvo hasta 1937. En 1937, el rendimiento se redujo a 360 kilogramos por hectárea, lo que representa una 40% de disminución con respecto al rendimiento de 1924.

A partir de 1950, el rendimiento comenzó a aumentar gradualmente. En 2000, el rendimiento alcanzó los 1,000 kilogramos por hectárea, lo que representa un 66% de aumento con respecto al rendimiento de 1950. En la última década, el rendimiento ha aumentado aún más. En 2020, el rendimiento alcanzó los 1,500 kilogramos por hectárea, lo que representa un 50% de aumento con respecto al rendimiento de 2000.

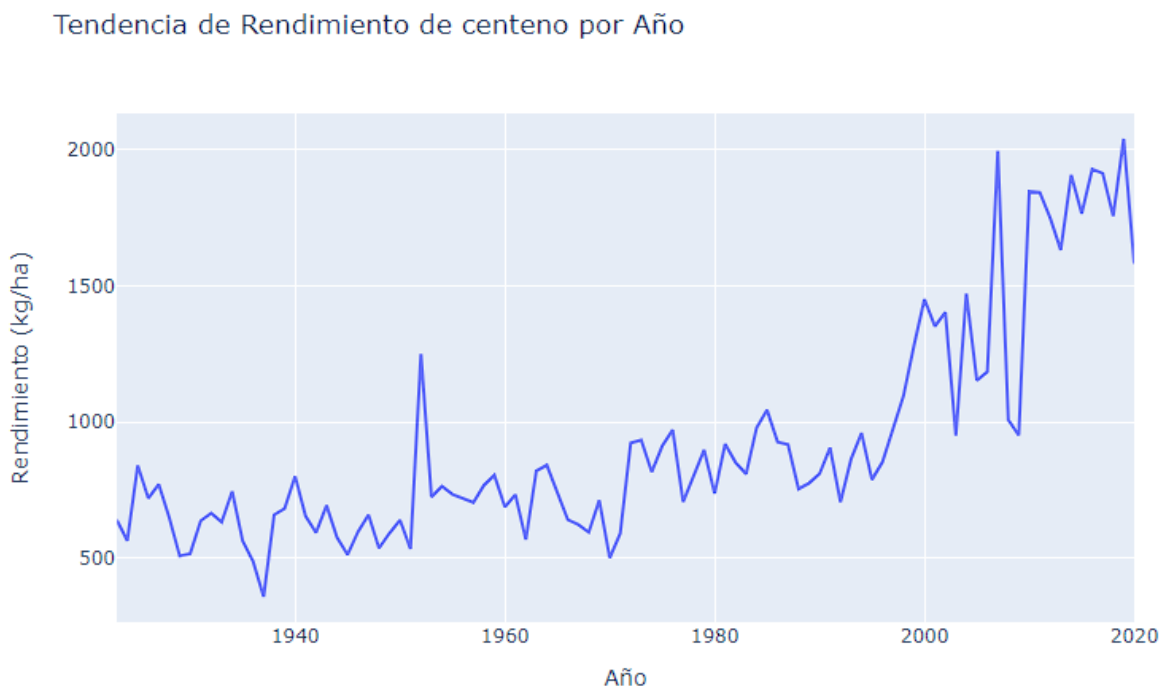

Figura 12: Rendimiento de centeno desde 1923 hasta 2020(kg. Cosechado por hectárea).

3.7 Conclusiones

Los análisis realizados sobre la producción y el rendimiento de los granos en Argentina muestran una serie de tendencias y factores que explican su evolución.

Tendencias generales

Aumento de la superficie destinada al cultivo de soja y maíz. La superficie destinada al cultivo de soja y maíz ha aumentado significativamente en las últimas décadas. Este aumento se puede atribuir a una serie de factores, entre los que se incluyen:

- El aumento de la demanda mundial de estos granos. La demanda mundial de soja y maíz ha aumentado significativamente en las últimas décadas, debido al crecimiento de la población y al aumento de la demanda de alimentos para animales.
- Los precios elevados de estos granos. Los precios de la soja y el maíz han sido históricamente elevados, lo que los ha hecho atractivos para los productores agrícolas.
- Las políticas gubernamentales de apoyo a la producción de soja y maíz. El gobierno argentino ha implementado una serie de políticas de apoyo a la producción de soja y maíz, como el subsidio a la semilla y a los fertilizantes.

Disminución de la superficie destinada al cultivo de centeno y girasol. La superficie destinada al cultivo de centeno y girasol ha disminuido significativamente en las últimas décadas. Este descenso se puede atribuir a:

- La disminución de la demanda mundial de estos granos. La demanda mundial de centeno y girasol ha disminuido significativamente en las últimas décadas, debido al aumento de la producción de otros granos, como la soja y el maíz.

- La competencia de otros cultivos. Otros cultivos, como la soja y el maíz, se han vuelto más competitivos que el centeno y el girasol, debido a su mayor rentabilidad.

Aumento del rendimiento de todos los granos. El rendimiento de todos los granos ha aumentado significativamente en las últimas décadas. Este aumento puede ser causado por:

- El desarrollo de nuevas tecnologías agrícolas. El desarrollo de nuevas tecnologías agrícolas, como las semillas mejoradas y la agricultura de precisión, ha permitido a los productores agrícolas aumentar la productividad de sus cultivos.
- La expansión de la frontera agrícola. La expansión de la frontera agrícola hacia zonas con condiciones climáticas favorables para el cultivo de granos ha permitido a los productores aumentar el rendimiento de sus cultivos.

Posibles factores que explican las decisiones de los productores

Las decisiones de los productores agrícolas sobre qué cultivos cultivar se basan en una serie de factores que pueden ser:

- La demanda y el precio de los granos. Los productores agrícolas cultivan los granos que tienen una alta demanda y un precio elevado.
- Los costos de producción. Los productores agrícolas consideran los costos de producción de los diferentes cultivos a la hora de tomar sus decisiones.
- Las políticas gubernamentales. Las políticas gubernamentales, como los subsidios a la producción, pueden influir en las decisiones de los productores agrícolas.

En el caso de Argentina, los factores que parecen haber influido en la decisión de los productores de aumentar la superficie destinada al cultivo de soja y maíz son:

- La alta demanda mundial de soja y maíz. La soja y el maíz son dos de los granos más demandados en el mundo, lo que ha impulsado los precios de estos granos.
- Los elevados precios de soja y maíz. Los precios de la soja y el maíz han sido históricamente elevados, lo que ha hecho que estos cultivos sean muy rentables para los productores agrícolas.
- Las políticas gubernamentales de apoyo a la producción de soja y maíz. El gobierno argentino ha implementado una serie de políticas de apoyo a la producción de soja y maíz, lo que ha reducido los costos de producción de estos cultivos.

En el caso de los productores de centeno y girasol, los factores que parecen haber influido en la decisión de disminuir la superficie destinada a estos cultivos son:

- La disminución de la demanda mundial de centeno y girasol. La demanda mundial de centeno y girasol ha disminuido significativamente en las últimas décadas, lo que ha reducido los precios de estos granos.
- La competencia de otros cultivos. Otros cultivos, como la soja y el maíz, se han vuelto más competitivos que el centeno y el girasol, debido a su mayor rentabilidad.

En conclusión, los análisis realizados sobre la producción y el rendimiento de los granos en Argentina muestran que la evolución de estos se debe a una serie de factores, entre los que se incluyen la demanda mundial, eventos climáticos, los precios, los costos de producción y las políticas gubernamentales.

4. Análisis exploratorio de trigo y soja.

En este análisis, nos sumergimos en la estimación de la producción de trigo y soja en Argentina, utilizando un conjunto de datos que abarca la producción de todos los cultivos, desglosado por provincia, departamento y otras características en todo el territorio argentino.

Primero, cargamos el conjunto de datos desde un archivo Excel y observamos las primeras filas para evaluar la organización. Notamos que la estructura del conjunto de datos inicial es inadecuada, separada por comas sin una estructura apta para el análisis y manipulación de datos.

Luego, aplicamos una función ‘organizar_datos’ para reorganizar y estructurar el conjunto de datos de manera que sea más manipulable. El DataFrame resultante se presenta para su visualización.

Posteriormente, exploramos las dimensiones del conjunto de datos, observando que tiene 153,889 filas y 12 columnas.

A continuación, verificamos los tipos de datos de las columnas y notamos que todas son de tipo 'Object' ('O'). Esto indica la necesidad de manipulación previa antes del análisis.

Dada la naturaleza de los datos, realizamos una manipulación de los tipos de datos para adecuarlos a los análisis correspondientes. Convertimos las columnas relevantes ('produccion', 'rendimiento', 'sup_cosechada', 'sup_sembrada') a tipos numéricos.

Adicionalmente, implementamos un proceso de saneado de datos para abordar posibles valores nulos ('NaN') o datos no deseados que pudieran interferir con los análisis.

Finalmente, convertimos las columnas numéricas a tipos de datos enteros para una representación más precisa y eficiente en los cálculos subsiguientes.

4.1 Selección y Filtrado de Columnas para el Análisis

En esta etapa, nos enfocamos en seleccionar las columnas clave que serán relevantes para nuestro análisis. Creamos un nuevo DataFrame (df_estimacion) que incluye solo las columnas que nos interesan, como 'id_provincia', 'provincia', 'departamento', 'id_cultivo', 'campania', 'cultivo', 'sup_sembrada', 'sup_cosechada', 'produccion' y 'rendimiento'.

Luego, inspeccionamos las columnas del nuevo DataFrame y verificamos los tipos de datos, confirmando que las columnas relevantes ahora son de tipo entero, lo cual es adecuado para manipulación y análisis.

Después, exploramos los valores únicos dentro de las columnas 'cultivo' y 'provincia' para identificar los cultivos y provincias disponibles para el análisis. Confirmamos que tenemos una lista completa de todas las provincias argentinas.

Por último, realizamos un filtrado específico para el cultivo de trigo y soja, respectivamente. Creamos DataFrames (df_filtrado_trigo y df_filtrado_soja) que contienen solo las filas relacionadas con los cultivos de trigo y soja. Esto nos permite centrarnos en el análisis de estas dos cosechas a nivel nacional.

5. Creación de modelos kmeans de sklearn para agrupar según el nivel de producción de las diferentes provincias.

5.1 Trigo producción por departamento.

5.1.1 Selección de características relevantes

Identificamos y seleccionamos las características pertinentes para nuestro análisis, 'produccion', 'rendimiento', 'sup_sembrada', 'sup_cosechada' considerándolas como principales para realizar el entrenamiento de los modelos KMeans.

5.1.2 Creación del Modelo KMeans de SKLearn

En esta etapa del proyecto, avanzamos en la creación y aplicación del modelo de agrupamiento KMeans utilizando la implementación proporcionada por la librería SKLearn. Nos propusimos identificar el número más adecuado de clusters para nuestro modelo de KMeans, utilizando una técnica conocida como el "Método del Codo" (Elbow Method). El objetivo es encontrar el punto en el que la suma de los errores cuadrados (SSE) disminuye significativamente, indicando el número óptimo de clusters.

5.1.3 Aplicación del Método del Codo

Se empleó el algoritmo KMeans del módulo sklearn.cluster para realizar agrupamientos en el conjunto de datos. Iteramos sobre un rango de posibles números de clusters (k) y calculamos la suma de errores cuadrados (SSE) para cada k. Los resultados se visualizaron en un gráfico, donde se observa la relación entre el número de clusters y la SSE.

Este proceso nos permitió identificar un punto en la curva donde la SSE comienza a estabilizarse, indicando el número óptimo de clusters. En el ejemplo proporcionado ver Figura 13, la curva del codo sugirió que el número ideal de clusters para nuestro conjunto de datos es 4.

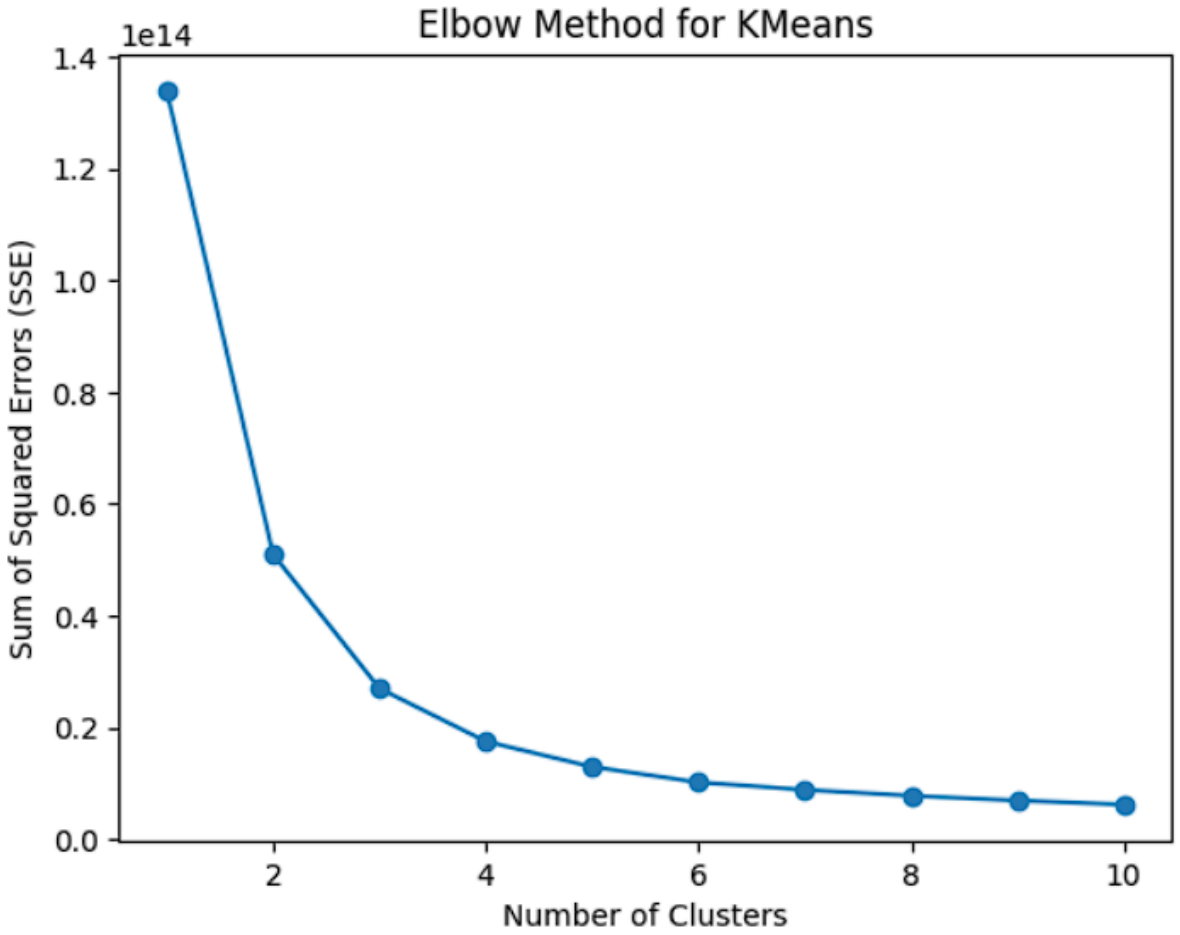

Figura 13: Curva del codo, ayuda a determinar el número de Clusters a utilizar.

Este análisis es crucial para determinar el nivel de partición más apropiado en los datos, proporcionando información valiosa para la fase siguiente del proyecto, donde se implementarán y evaluarán los modelos de agrupamiento. En las siguientes secciones, profundizaremos en la aplicación y análisis de los resultados de KMeans con el número óptimo de clusters identificado

5.1.4 Implementación del Modelo KMeans con Número Óptimo de Clusters

Con el número óptimo de clusters identificado como 3 o 4 mediante el Método del Codo, procedimos a la creación y aplicación del modelo KMeans a nuestro conjunto de datos.

5.1.5 Instanciación del Modelo KMeans

Creamos una instancia del modelo KMeans de la librería SKLearn, especificando el número óptimo de clusters identificado anteriormente, que en este caso fue 4. La inicialización se llevó a cabo con una semilla aleatoria (random_state=42) para garantizar reproducibilidad.

5.1.6 Ajuste del Modelo a los Datos

Aplicamos el modelo KMeans al conjunto de datos utilizando la función fit_predict, lo que nos permitió asignar cada punto de datos a un cluster particular. La información de los clusters resultantes se incorporó al DataFrame original para futuros análisis.

5.1.7 Visualización de los Resultados

Creación de visualizaciones para entender y comunicar efectivamente los resultados del modelo KMeans. Utilizamos un gráfico de dispersión para representar la agrupación de datos en función de las características 'sup_sembrada', 'sup_cosechada', 'produccion' y 'rendimiento', pudiéndose observar la relación entre estos (más superficie cosechada equivale a mayor producción) ver figura 14. Y luego creamos diferentes gráficos utilizando la función pairplot para ver estas relaciones y poder apreciar que está indicando cada cluster, para luego crear etiquetas descriptivas según el nivel de producción.

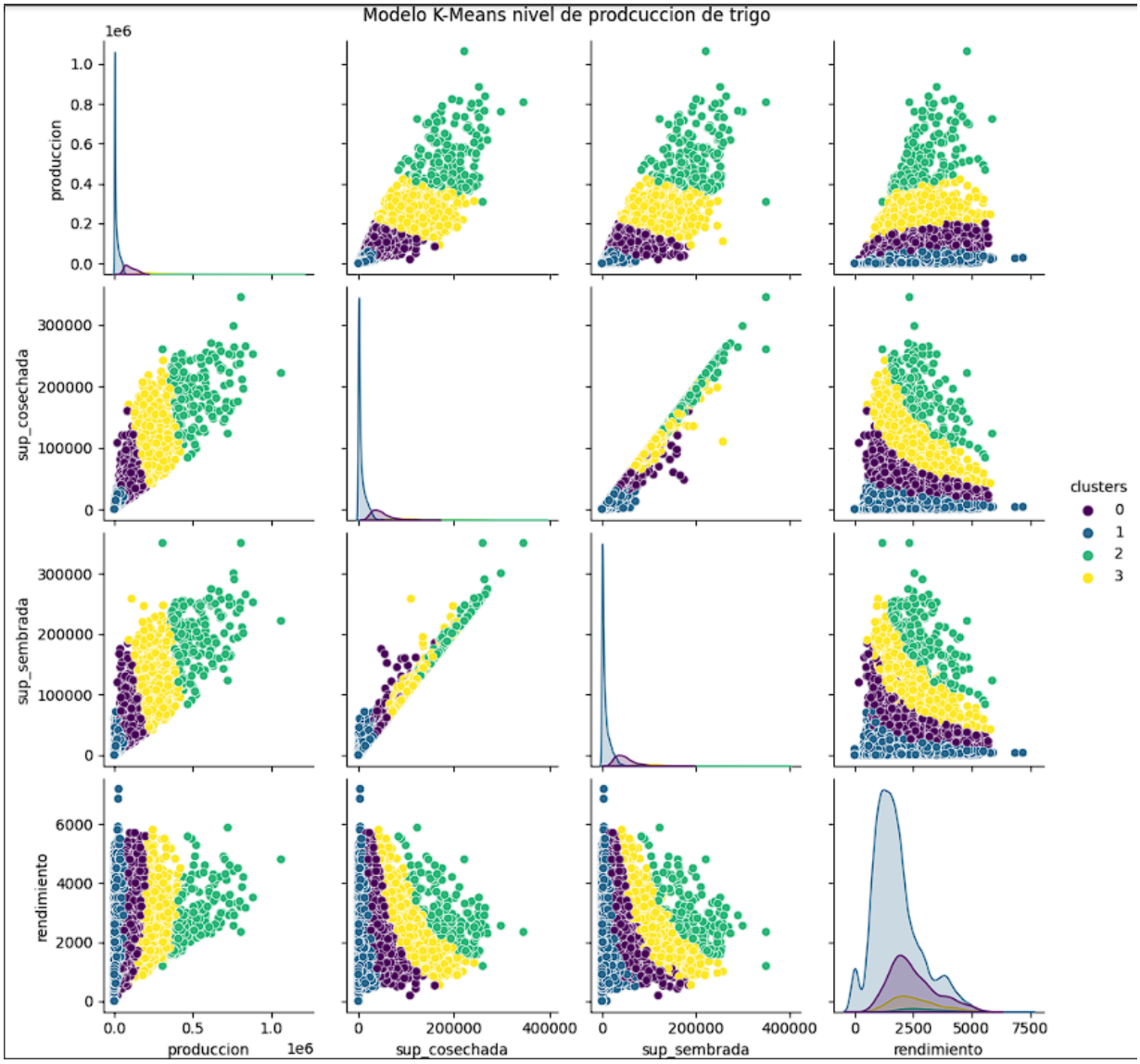

Figura 14: Gráfico con pairplot, relaciones varias entre variables modelo KMeans.

5.1.8 Etiquetado y visualización de Niveles de Producción

Para facilitar la interpretación de los resultados, asignamos etiquetas descriptivas a cada cluster. En este caso, se utilizaron las etiquetas 'bajo', 'medio', 'alto', 'muy alto' para representar los clusters 1, 0, 3 y 2 respectivamente. Estas etiquetas se añadieron como una nueva columna llamada 'nivel_produccion' al DataFrame. Luego visualizamos nuevamente utilizando pairplot para apreciar cada cluster con su nivel de producción correspondiente ver figura 15.

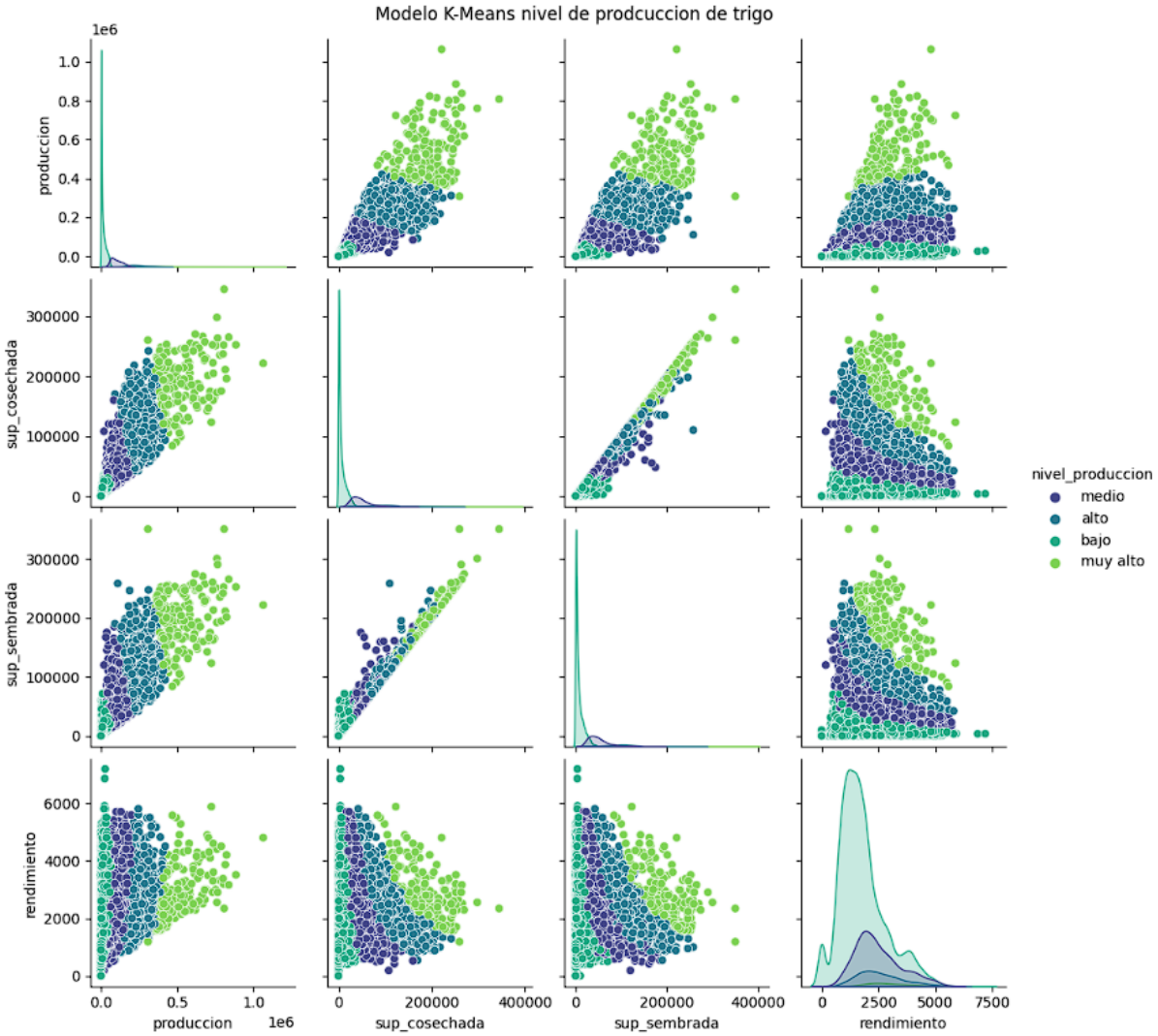

Figura 15: Gráfico con pairplot, relaciones varias entre variables modelo KMeans.

5.2 Trigo producción por provincia.

5.2.1 Preparado de datos

Primero reutilizamos el DataFrame filtrado en el punto anterior por cultivo de Trigo y a partir de ese DF reducimos las columnas a las que son de nuestro interés para el modelo. Luego agrupamos el nuevo DataFrame por provincias y hacemos una sumatoria de todas las columnas numéricas y creamos un nuevo DataFrame llamado ‘df_trigo_agrupado_por_provincia’, ver figura 16.

df_trigo_agrupado_por_provincia

	provincia	produccion	rendimiento	sup_cosechada	sup_sembrada
0	Buenos Aires	365426402	13254423	154392727	160939168
1	Catamarca	835464	189534	519630	556530
2	Chaco	3920075	1116852	2811093	3303576
3	Chubut	3890	27741	4000	9525
4	Cordoba	84027691	1738993	38653094	42609612
5	Corrientes	81276	126333	44988	48845
6	Entre Rios	30757549	1662576	12289655	12960205
7	Formosa	103501	123081	74710	89320
8	Jujuy	148343	387298	78265	87019
9	La Pampa	32769363	1332422	20498705	23294435
10	La Rioja	344	15724	346	684
11	Misiones	80	9583	100	100
12	Neuquen	8885	125152	11925	13380
13	Rio Negro	144556	126172	115167	145235
14	Salta	3310512	659690	2533624	2738263
15	San Juan	16020	71306	9454	10114
16	San Luis	628083	384182	332035	429155
17	Santa Cruz	150	7000	150	1809
18	Santa Fe	101126485	2164131	42321351	44265442
19	Santiago del Estero	10035074	1117706	5493381	6288526
20	Tucuman	4041767	445399	3245871	3652880

Figura 16: DataFrame agrupado por provincias ‘df_trigo_agrupado_por_provincia’.

5.2.2 Selección de características relevantes

Identificamos y seleccionamos las características pertinentes para nuestro análisis, 'produccion', 'rendimiento', 'sup_sembrada', 'sup_cosechada' considerándolas como principales para realizar el entrenamiento de los modelos KMeans.

5.2.3 Busqueda del K (Número de clusters) Óptimo para nuestro modelo

Iteramos sobre un rango de posibles K y calculamos la suma de errores cuadrados (SSE) para cada K. Los resultados se visualizan en el siguiente gráfico, donde se observa la relación entre el número de clusters y la SSE. Este proceso nos permite identificar un punto en la curva donde la SSE comienza a estabilizarse, indicando el número óptimo de clusters. Conocido como Elbow Method o Curva/Método del Codo en español, ver figura 17.

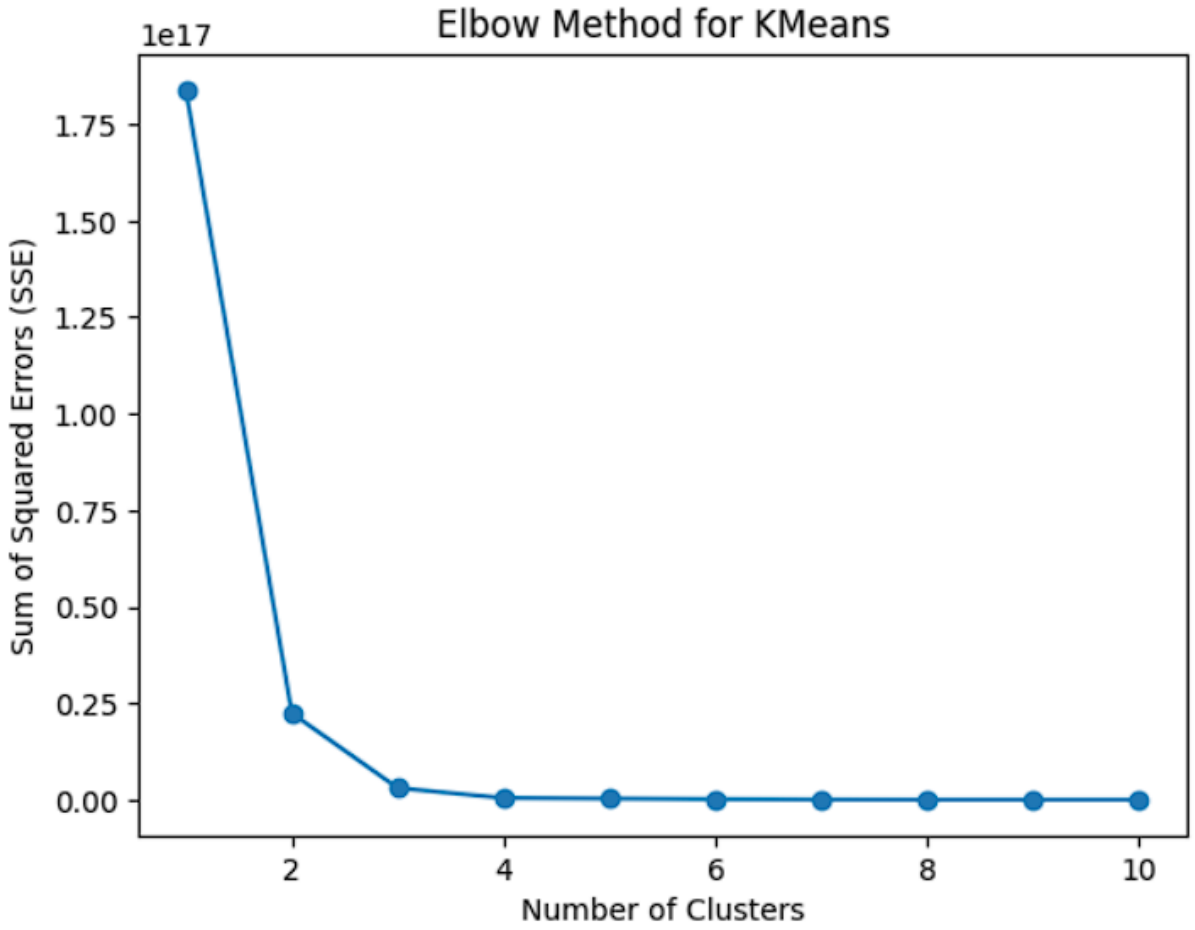

Figura 17: Curva del codo, ayuda a determinar el número de Clusters a utilizar.

5.2.4 Instanciación del Modelo KMeans

Creamos una instancia del modelo KMeans de la librería SKLearn, especificando el número óptimo de clusters identificado anteriormente, que en este caso fue 4. La inicialización se llevó a cabo con una semilla aleatoria (random_state=42) para garantizar reproducibilidad.

5.2.5 Ajuste del Modelo a los Datos

Aplicamos el modelo KMeans al conjunto de datos utilizando la función fit_predict, lo que nos permitió asignar cada punto de datos a un cluster particular. La información de los clusters resultantes se incorporó al DataFrame original para futuros análisis.

5.2.6 Visualización de los Resultados

Creación de visualizaciones para entender y comunicar efectivamente los resultados del modelo KMeans. Utilizamos un gráfico de dispersión para representar la agrupación de datos en función de las características 'sup_sembrada', 'sup_cosechada', 'produccion' y 'rendimiento', pudiéndose observar la relación entre estos (más superficie cosechada equivale a mayor

producción) ver figura 18. Y luego creamos diferentes gráficos utilizando la función pairplot para ver estas relaciones y poder apreciar que está indicando cada cluster, para luego crear etiquetas descriptivas según el nivel de producción.

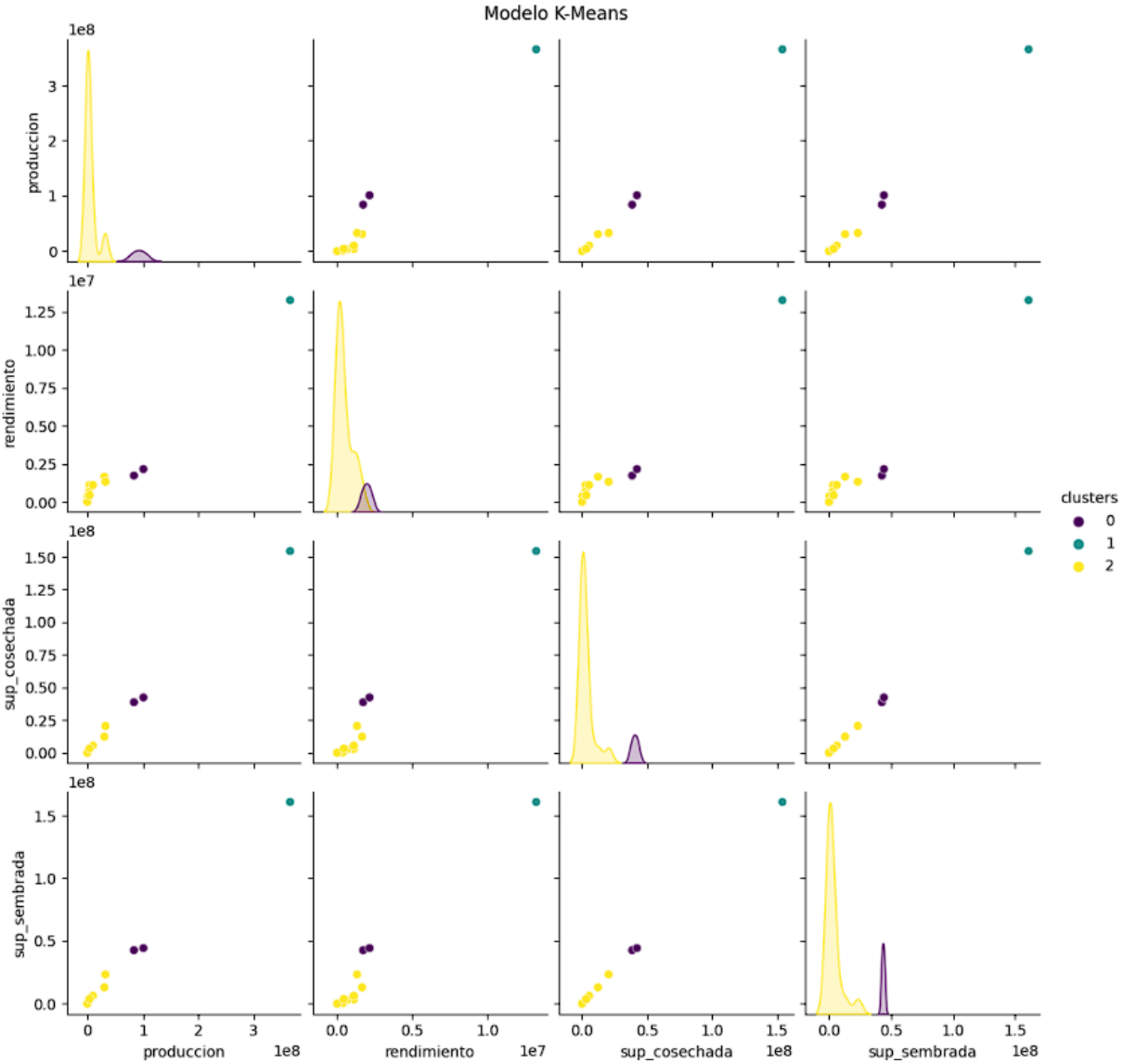

Figura 18: Gráfico con pairplot, relaciones varias entre variables modelo KMeans.

5.2.7 Etiquetado de Niveles de Producción.

Para facilitar la interpretación de los resultados, asignamos etiquetas descriptivas a cada cluster. En este caso, se utilizaron las etiquetas 'Medio Bajo', 'Medio Alto' y 'Muy Alto' para representar los clusters 2, 0 y 1 respectivamente. Estas etiquetas se añadieron como una nueva columna llamada 'nivel_produccion' al DataFrame utilizado para este modelo.

5.2.8 Visualización Interactiva de los Resultados con etiquetas discriminadas por nivel de producción.

En esta ocasión utilizamos el gráfico scatter_matrix de Plotly, que guarda similitudes con el gráfico pairplot de Seaborn, pero con ventajas ya que estos gráficos son interactivos para el

usuario, ofreciendo una mejor experiencia y un detalle más profundo del agrupamiento realizado por nuestro modelo KMeans, como, por ejemplo, al hacer hover por cada punto discrimina el departamento, su producción total, su nivel de producción, la superficie cosechada y la superficie sembrada, también se puede ver la clasificación por etiquetas, ver figura 19.

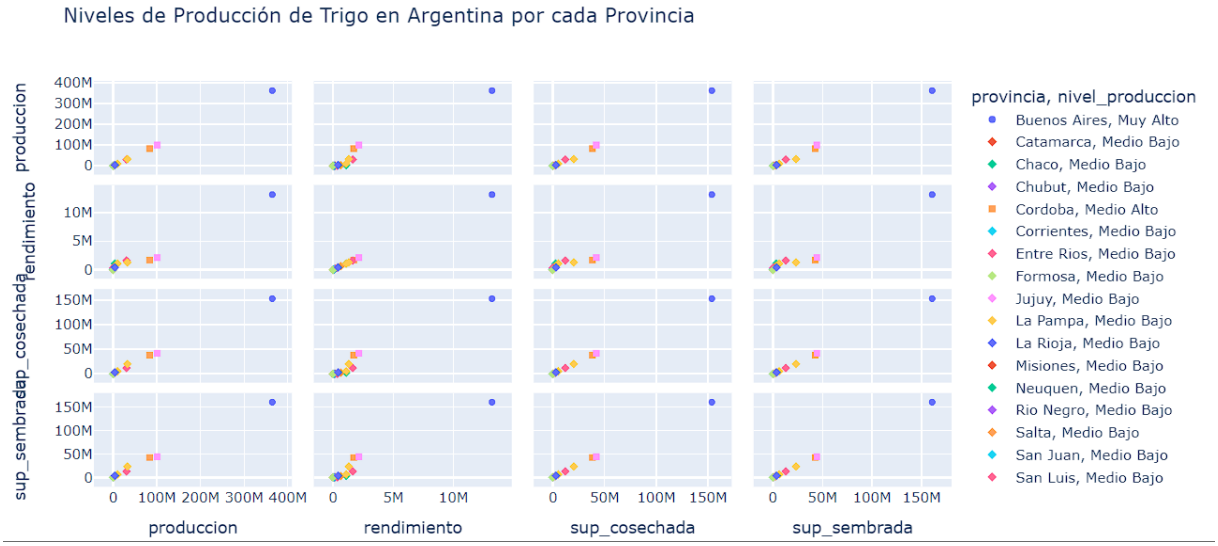

Figura 19: Gráficos de dispersión con la función scatter_matrix de la librería Plotly.

5.3 SOJA PRODUCCIÓN POR DEPARTAMENTO.

5.3.1 Reutilización de datos y selección de características relevantes

Primero reutilizamos el DataFrame que tenemos filtrado que solo contiene el cultivo de soja (`df_filtrado_soja`). Identificamos y seleccionamos las características pertinentes para nuestro análisis, 'produccion', 'rendimiento', 'sup_sembrada', 'sup_cosechada' considerándolas como principales para realizar el entrenamiento de los modelos KMeans.

5.3.2 Creación del Modelo KMeans de SKLearn

En esta etapa del proyecto, avanzamos en la creación y aplicación del modelo de agrupamiento KMeans utilizando la implementación proporcionada por la librería SKLearn. Nos propusimos identificar el número más adecuado de clusters para nuestro modelo de KMeans, utilizando una técnica conocida como el "Método del Codo" (Elbow Method). El objetivo es encontrar el punto en el que la suma de los errores cuadrados (SSE) disminuye significativamente, indicando el número óptimo de clusters.

5.3.3 Aplicación del Método del Codo

Se empleó el algoritmo KMeans del módulo sklearn.cluster para realizar agrupamientos en el conjunto de datos. Iteramos sobre un rango de posibles números de clusters (k) y calculamos la suma de errores cuadrados (SSE) para cada k. Los resultados se visualizaron en un gráfico, donde se observa la relación entre el número de clusters y la SSE.

Este proceso nos permitió identificar un punto en la curva donde la SSE comienza a estabilizarse, indicando el número óptimo de clusters. En el ejemplo proporcionado ver Figura 20, la curva del codo sugirió que el número ideal de clusters para nuestro conjunto de datos es 4.

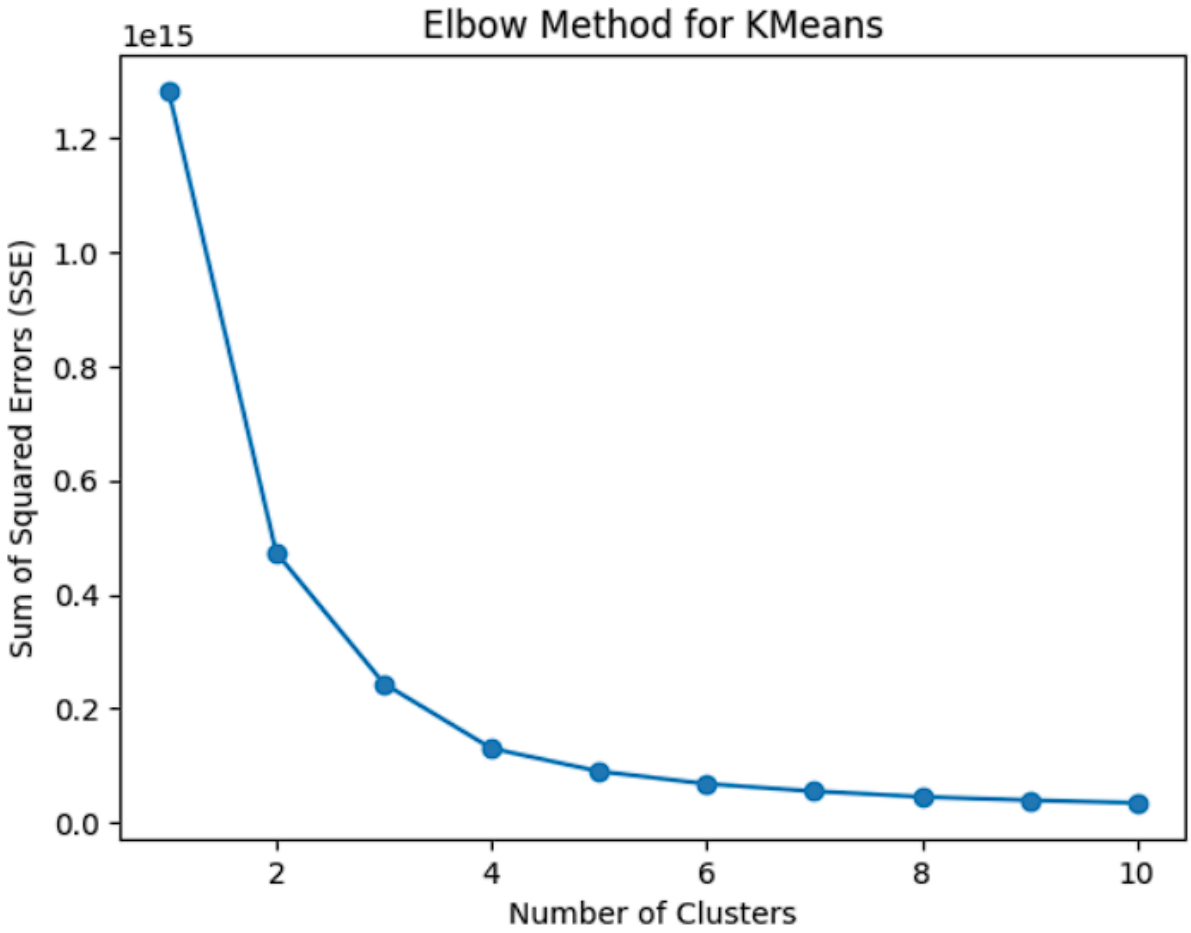

Figura 13: Curva del codo, ayuda a determinar el número de Clusters a utilizar.

Este análisis es crucial para determinar el nivel de partición más apropiado en los datos, proporcionando información valiosa para la fase siguiente del proyecto, donde se implementarán y evaluarán los modelos de agrupamiento. En las siguientes secciones, profundizaremos en la aplicación y análisis de los resultados de KMeans con el número óptimo de clusters identificado

5.3.4 Implementación del Modelo KMeans con Número Óptimo de Clusters

Con el número óptimo de clusters identificado como 3 o 4 mediante el Método del Codo, procedimos a la creación y aplicación del modelo KMeans a nuestro conjunto de datos.

5.3.5 Instanciación del Modelo KMeans

Creamos una instancia del modelo KMeans de la librería SKLearn, especificando el número óptimo de clusters identificado anteriormente, que en este caso fue 4. La inicialización se llevó a cabo con una semilla aleatoria (random_state=42) para garantizar reproducibilidad.

5.3.6 Ajuste del Modelo a los Datos

Aplicamos el modelo KMeans al conjunto de datos utilizando la función fit_predict, lo que nos permitió asignar cada punto de datos a un cluster particular. La información de los clusters resultantes se incorporó al DataFrame original para futuros análisis.

5.3.7 Visualización de los Resultados

Creación de visualizaciones para entender y comunicar efectivamente los resultados del modelo KMeans. Utilizamos un gráfico de dispersión para representar la agrupación de datos en función de las características 'sup_sembrada', 'sup_cosechada', 'produccion' y 'rendimiento', pudiéndose observar la relación entre estos (más superficie cosechada equivale a mayor producción) ver figura 21. Y luego creamos diferentes gráficos utilizando la función pairplot para ver estas relaciones y poder apreciar que está indicando cada cluster, para luego crear etiquetas descriptivas según el nivel de producción.

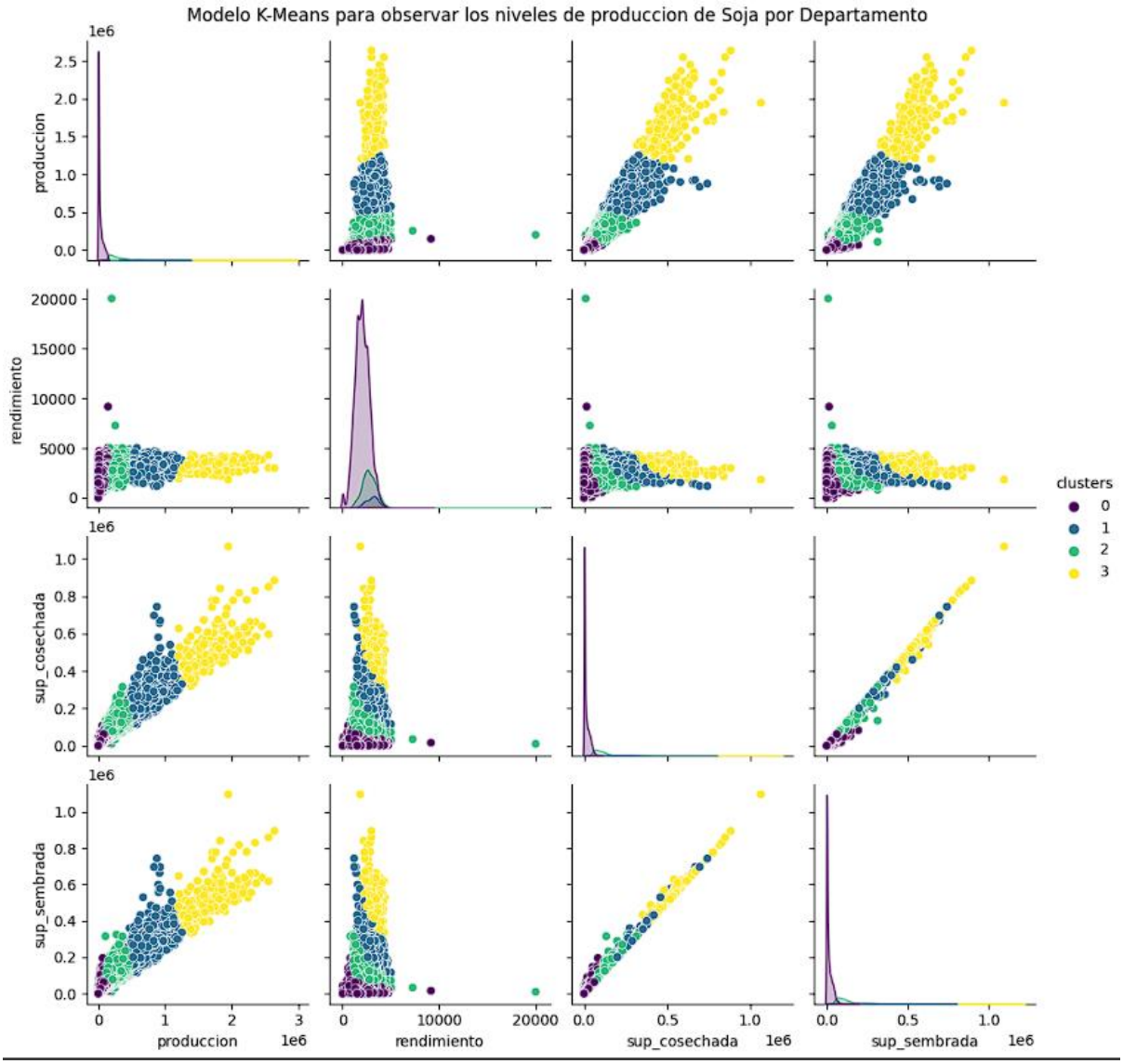

Figura 21: Gráfico con pairplot, relaciones varias entre variables modelo KMeans.

5.3.8 Etiquetado y visualización de Niveles de Producción

Para facilitar la interpretación de los resultados, asignamos etiquetas descriptivas a cada cluster. En este caso, se utilizaron las etiquetas 'bajo', 'medio', 'alto', 'muy alto' para representar los clusters 1, 0, 3 y 2 respectivamente. Estas etiquetas se añadieron como una nueva columna llamada 'nivel_produccion' al DataFrame. Luego visualizamos nuevamente utilizando pairplot para apreciar cada cluster con su nivel de producción correspondiente ver figura 22.

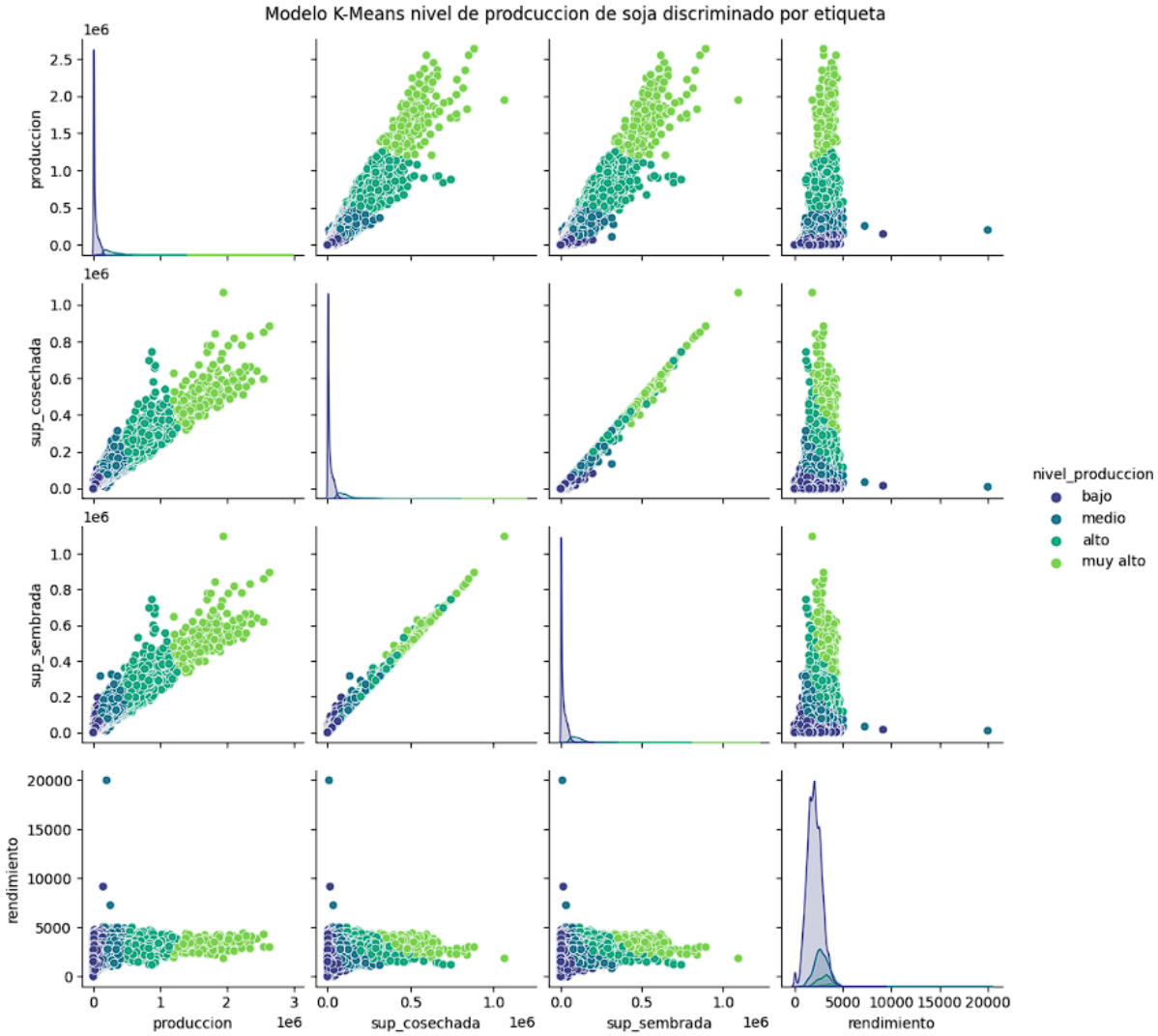

Figura 22: Gráfico con pairplot, relaciones varias entre variables modelo KMeans.

5.4 Sojaproducción por provincia.

5.4.1 Preparado de datos

Primero reutilizamos el DataFrame filtrado en el punto anterior por cultivo de Soja y a partir de ese DataFrame reducimos las columnas a las que son de nuestro interés para el modelo. Luego agrupamos el nuevo dataframe por provincias y hacemos una sumatoria de todas las columnas numéricas y creamos un nuevo DataFrame llamado 'df_agrupado_por_provincia_soja', ver figura 23.

df_agrupado_por_provincia_soja

	provincia	produccion	rendimiento	sup_cosechada	sup_sembrada
0	Buenos Aires	643341842	18407392	240906088	247890023
1	Catamarca	4783545	406339	1941102	1973012
2	Chaco	55791812	3256948	27563405	29823015
3	Cordoba	623429710	4518096	232517732	237548212
4	Corrientes	1819576	947533	1122050	1219463
5	Entre Rios	131503951	2905792	57044647	59266194
6	Formosa	1240072	445215	589725	687440
7	Jujuy	862986	834243	358505	371893
8	La Pampa	38247391	1768692	15708999	16445452
9	Mendoza	0	0	0	35
10	Misiones	940336	1194206	602750	673958
11	Salta	49570125	1919476	21763399	22760225
12	San Luis	19084762	740326	8574794	8697724
13	Santa Fe	521767435	4376546	183480570	188332520
14	Santiago del Estero	99468113	3183392	39928607	42268661
15	Tucuman	28354636	2057423	11969022	12189238

Figura 22: DataFrame agrupado por provincias ‘df_agreupado_por_provincia_soja’.

5.4.2 Selección de características relevantes

Identificamos y seleccionamos las características pertinentes para nuestro análisis, 'produccion', 'rendimiento', 'sup_sembrada', 'sup_cosechada' considerándolas como principales para realizar el entrenamiento de los modelos KMeans.

5.4.3 Busqueda del K (Número de clusters) Óptimo para nuestro modelo

Iteramos sobre un rango de posibles K y calculamos la suma de errores cuadrados (SSE) para cada K. Los resultados se visualizan en el siguiente gráfico, donde se observa la relación entre el número de clusters y la SSE. Este proceso nos permite identificar un punto en la curva donde la SSE comienza a estabilizarse, indicando el número óptimo de clusters. Conocido como Elbow Method o Curva/Método del Codo en español, ver figura 23.

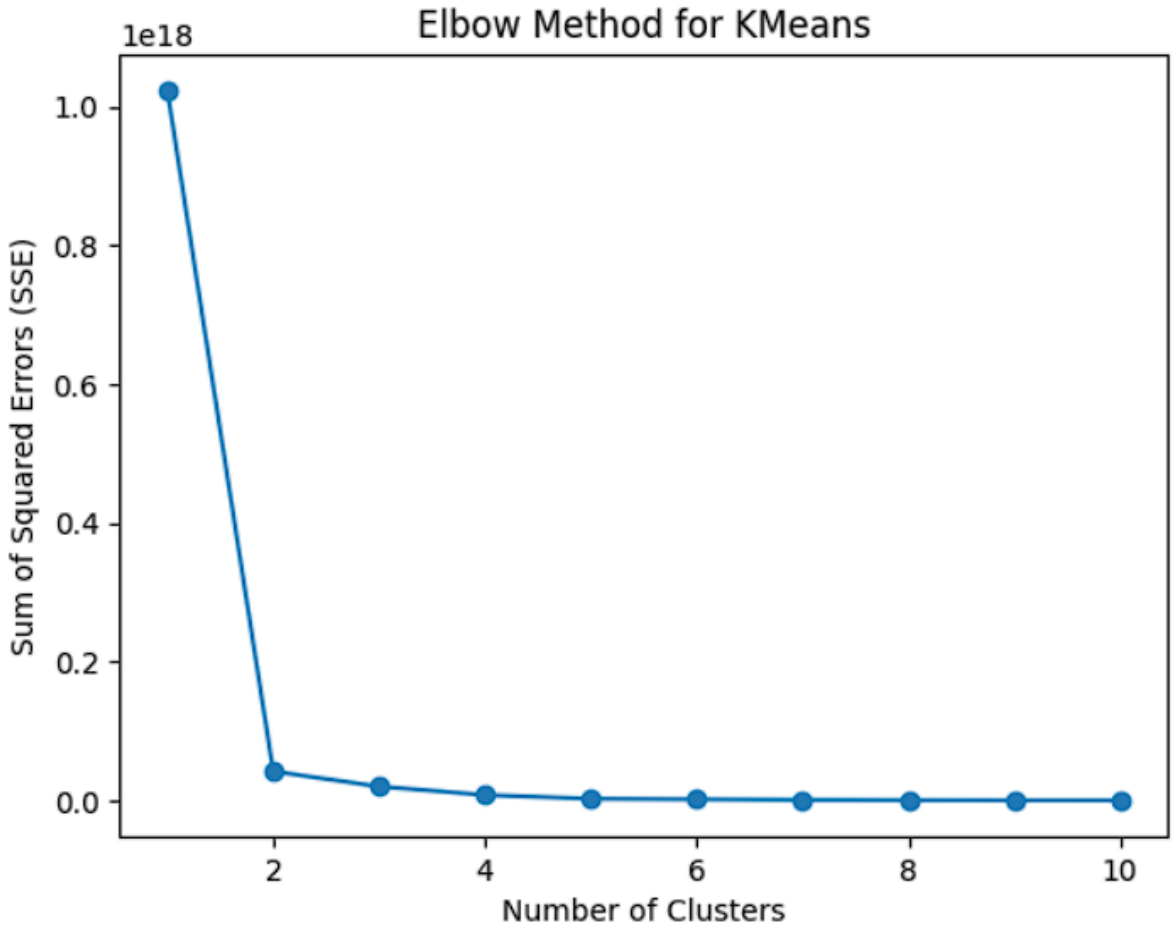

Figura 24: Curva del codo, ayuda a determinar el número de Clusters a utilizar.

5.4.4 Instanciación del Modelo KMeans

Creamos una instancia del modelo KMeans de la librería SKLearn, especificando el número óptimo de clusters identificado anteriormente, que en este caso fue 2. La inicialización se llevó a cabo con una semilla aleatoria (random_state=42) para garantizar reproducibilidad.

5.4.5 Ajuste del Modelo a los Datos

Aplicamos el modelo KMeans al conjunto de datos utilizando la función fit_predict, lo que nos permitió asignar cada punto de datos a un cluster particular. La información de los clusters resultantes se incorporó al DataFrame original para futuros análisis.

5.4.6 Visualización de los Resultados

Creación de visualizaciones para entender y comunicar efectivamente los resultados del modelo KMeans. Utilizamos un gráfico de dispersión para representar la agrupación de datos en función de las características 'sup_sembrada', 'sup_cosechada', 'produccion' y 'rendimiento', pudiéndose observar la relación entre estos (más superficie cosechada equivale a mayor producción) ver figura 18. Y luego creamos diferentes gráficos utilizando la función pairplot para ver estas relaciones y poder apreciar que está indicando cada cluster, para luego crear etiquetas descriptivas según el nivel de producción.

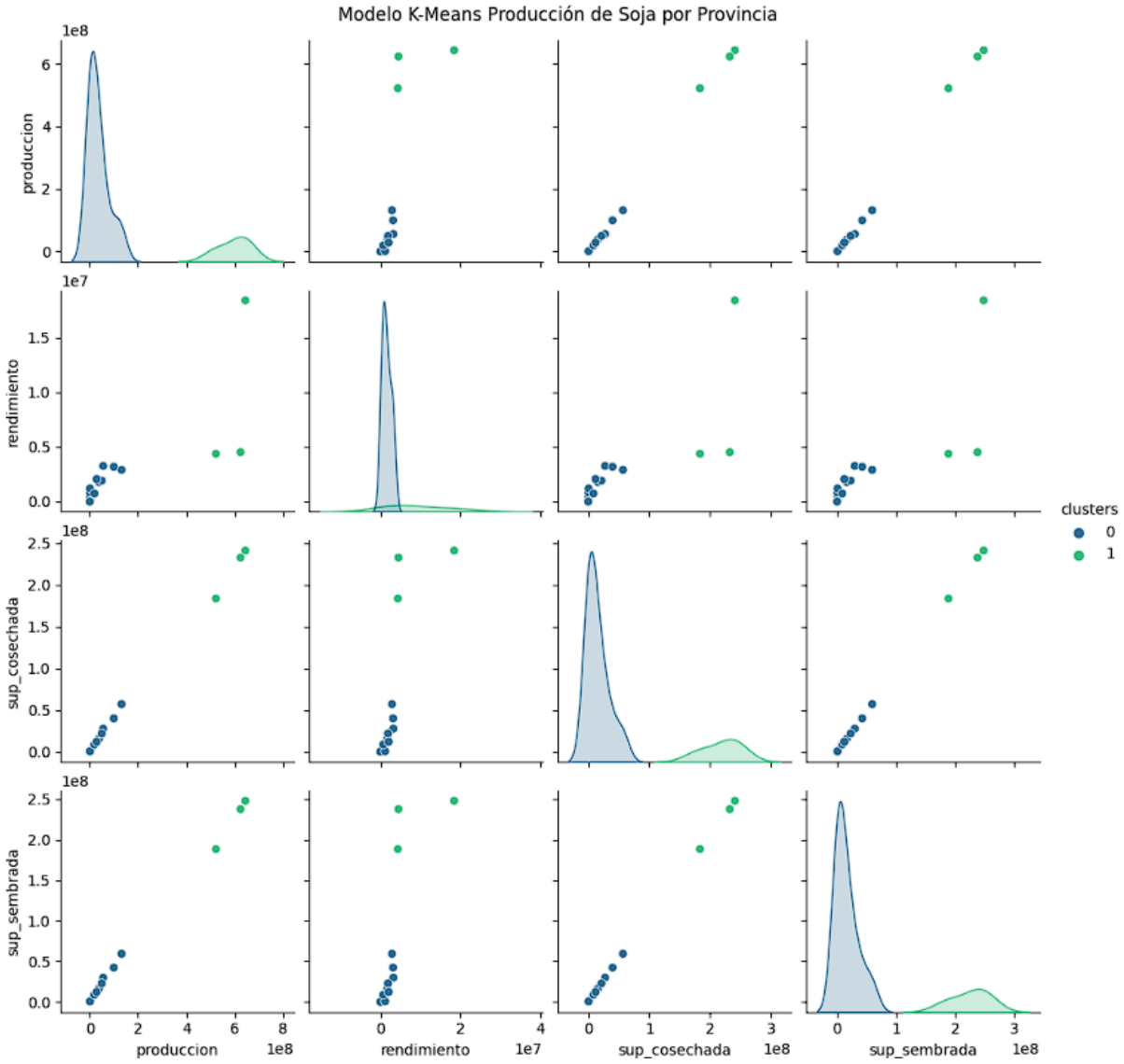

Figura 22: Gráfico con pairplot, relaciones varias entre variables modelo KMeans.

5.4.7 Etiquetado de Niveles de Producción.

Para facilitar la interpretación de los resultados, asignamos etiquetas descriptivas a cada cluster. En este caso, se utilizaron las etiquetas 'bajo' y 'alto' para representar los clusters 0 y 1 respectivamente. Estas etiquetas se añadieron como una nueva columna llamada 'nivel_produccion' al DataFrame utilizado para este modelo.

5.4.8 Visualización Interactiva de los Resultados con etiquetas discriminadas por nivel de producción.

En esta ocasión utilizamos el gráfico scatter_matrix de Plotly, que guarda similitudes con el gráfico pairplot de Seaborn, pero con ventajas ya que estos gráficos son interactivos para el usuario, ofreciendo una mejor experiencia y un detalle más profundo del agrupamiento realizado por nuestro modelo KMeans, como, por ejemplo, al hacer hover por cada punto discrimina el departamento, su producción total, su nivel de producción, la superficie

cosechada y la superficie sembrada, también se puede ver la clasificación por etiquetas, ver figura 23.

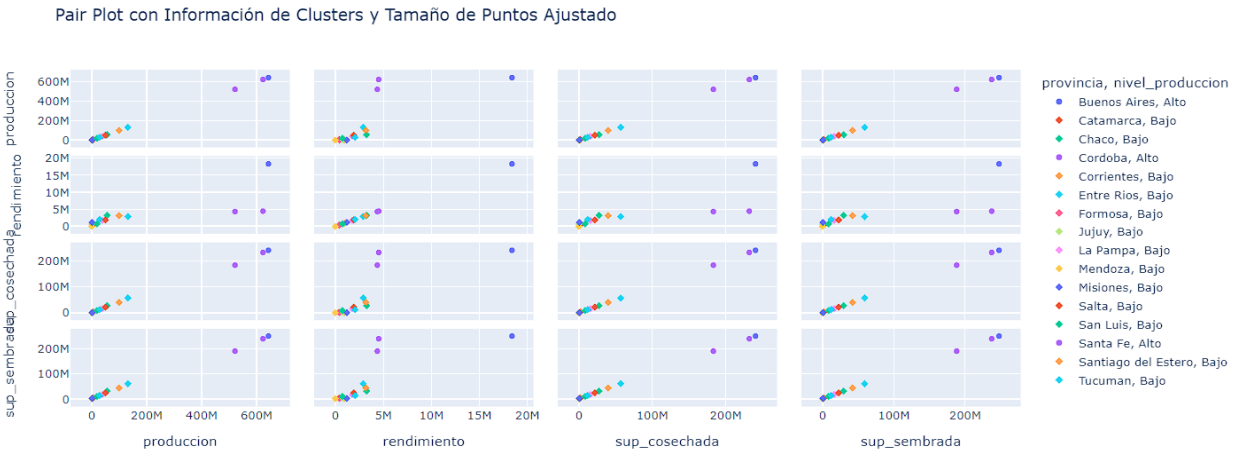

Figura 23: Gráficos de dispersión con la función scatter_matrix de la librería Plotly.

5.5 Conclusiones

Relación Entre Superficie Cosechada y Producción

Se puede observar claramente que a medida que la superficie cosechada aumenta, también lo hace la producción de soja en los diferentes departamentos. Este patrón sugiere que existe una relación positiva entre la cantidad de tierra dedicada a la cosecha y la producción obtenida.

Relación Entre Superficie Sembrada y Superficie Cosechada

El gráfico también revela una tendencia que indica que a medida que la superficie sembrada aumenta, la superficie cosechada tiende a seguir la misma dirección. Esto sugiere que, en general, a mayor superficie sembrada, se espera una mayor cantidad de tierra que finalmente se cosecha.

Por lo tanto, después del análisis visual de las relaciones en los gráficos respalda que la superficie dedicada a la cosecha y siembra tiene una influencia significativa en los niveles de producción tanto del Trigo como para la Soja.

Destaque Provincial

Adicionalmente, al examinar detenidamente los gráficos, se puede concluir que la provincia de Buenos Aires destaca por encima de todas en términos de producción, tanto para el Trigo como para la Soja. Siguiéndole en producción las provincias de Córdoba y Santa Fe respectivamente. Este patrón se repite consistentemente en ambos cultivos, sugiriendo una influencia regional significativa en los niveles de producción. Las demás provincias muestran niveles de producción inferiores en comparación con estas destacadas.

6. Reflexión Final

Resumen de la Primera Parte del Proyecto

Los análisis sobre la producción y rendimiento de granos en Argentina revelan tendencias clave y factores determinantes de su evolución. Se destaca el notorio aumento en la superficie destinada a la siembra de soja y maíz, influenciado por la creciente demanda global, precios históricamente elevados, y políticas gubernamentales de apoyo. En contraste, la disminución de la superficie para centeno y girasol se atribuye a la baja demanda mundial y la competencia con cultivos más rentables como la soja y el maíz. Además, el incremento general del rendimiento de todos los granos se vincula con avances tecnológicos y la expansión de áreas agrícolas.

Reflexiones sobre las Decisiones de los Productores

Las decisiones de los productores agrícolas en Argentina están fuertemente influenciadas por la demanda y los precios de los granos, así como por los costos de producción y las políticas gubernamentales. Para soja y maíz, la alta demanda mundial y políticas de respaldo han impulsado su expansión. En cambio, productores de centeno y girasol se han visto afectados por la disminución de la demanda global y la creciente competencia. Este análisis subraya la interconexión de factores económicos, políticos y de mercado que dan forma a las decisiones agrícolas en el país.

Reflexiones de la Segunda Parte del Proyecto

La relación positiva entre la superficie cosechada y la producción de soja en diferentes departamentos es evidente. Este patrón indica que una mayor área dedicada a la cosecha se traduce en una producción más significativa. Asimismo, la relación entre la superficie sembrada y cosechada refleja que un aumento en la siembra generalmente conlleva a una cosecha proporcionalmente mayor. Estos hallazgos respaldan la importancia de la superficie dedicada a la cosecha y siembra en la determinación de los niveles de producción tanto para el trigo como para la soja.

Destaque Provincial

Al profundizar en los gráficos, surge una conclusión destacada: Buenos Aires sobresale como la provincia líder en producción, tanto para el trigo como para la soja. Córdoba y Santa Fe le siguen en términos de producción, consolidando un patrón regional influyente en los niveles de producción. Las demás provincias exhiben niveles de producción inferiores en comparación con estas destacadas. Este análisis regional proporciona una perspectiva valiosa para comprender las dinámicas específicas de cada provincia en el contexto agrícola.

En conjunto, estos análisis integran una visión detallada de la evolución y determinantes de la producción de granos en Argentina. La interacción compleja entre factores económicos, climáticos y políticos destaca la necesidad de un enfoque integral para comprender y anticipar los cambios en la agricultura del país.