

Trabajo 2: Procesamiento estadístico de datos

Juan David Garcia Zapata
CC: 1001660790

Julián Castaño Pineda
CC: 1000755372

Mauricio Alejandro Mazo Lopera



Introducción

El propósito de este segundo trabajo es familiarizarse con el uso de **R Markdown** para la creación de informes, y aplicar estos conocimientos en el análisis de una base de datos. El objetivo principal es extraer información estadística relevante de los datos disponibles. A continuación, se detallan los puntos que guiarán el análisis:

1. Contextualización de la Base de Datos

Se debe contextualizar la base de datos, explicando a qué se refiere y describiendo sus dimensiones, es decir, el número de filas y columnas. Además, se debe crear una tabla que contenga los nombres de las variables para facilitar su comprensión.

2. Selección de Variables y Análisis Descriptivo

Se debe seleccionar una o más variables que se consideren de mayor relevancia dentro de la base de datos. Luego, plantear un objetivo específico para el análisis de estas variables y realizar un análisis descriptivo detallado.

3. Identificación de Covariables

Se debe identificar un conjunto de covariables o variables explicativas que puedan tener una relación con las variables seleccionadas en el punto anterior (ítem 2). Estas covariables serán importantes para el análisis posterior.

4. Análisis de Covariables

Finalmente, se deben analizar las covariables seleccionadas mediante tablas, resúmenes estadísticos y/o gráficos. Este análisis debe tener en cuenta el tipo de cada covariable y cómo se relacionan con las variables seleccionadas en el ítem (2).

a) Contextualización de la Base de Datos

La base de datos “Rendimiento de Cultivos en Colombia por Año” se encuentra disponible públicamente a través de la plataforma **Datos Abiertos del Gobierno de Colombia**. Esta base proporciona información valiosa y detallada sobre el rendimiento de los cultivos agrícolas en diversas regiones del país, abarcando los años desde 2007 hasta 2018.

Contexto

El sector agrícola es un componente esencial de la economía colombiana, y el rendimiento de los cultivos es un indicador clave para evaluar la productividad de este sector. Los datos recopilados en esta base de datos provienen de fuentes oficiales del Departamento Administrativo Nacional de Estadística (DANE) y proporcionan una visión integral de las prácticas agrícolas en el país. Esta base cubre un amplio espectro de cultivos, que incluye tanto productos tradicionales como otros más recientes, lo que permite a los investigadores y tomadores de decisiones analizar las tendencias y variaciones en el rendimiento a lo largo del tiempo.

Dimensiones de la Base de Datos

La base de datos consta de **Número de filas:** 206068 y **Número de columnas:** 17, lo que implica que contiene una cantidad significativa de registros (uno por cada combinación de región y cultivo) a lo largo de los años estudiados. Esto ofrece un panorama muy detallado de cómo ha evolucionado el rendimiento de los cultivos en distintas regiones del país. La cantidad de variables (columnas) presentes en la base de datos también es considerable, lo que nos permite realizar análisis multifacéticos, considerando no solo los rendimientos, sino también factores relacionados, como la región geográfica, el tipo de cultivo, las variaciones estacionales, y más.

Table 1: Variables de la Base de Datos

CÓD. DEP.	CULTIVO	Producción (t)
DEPARTAMENTO	DESAGREGACIÓN REGIONAL Y/O SISTEMA PRODUCTIVO	Rendimiento (t/ha)
CÓD. MUN.	AÑO	ESTADO FISICO PRODUCCION
MUNICIPIO	PERIODO	NOMBRE CIENTIFICO
GRUPO DE CULTIVO	Área Sembrada (ha)	CICLO DE CULTIVO
SUBGRUPO DE CULTIVO	Área Cosechada (ha)	

La Tabla 1 detalla las variables que integran la base de datos. A continuación se describe cada una de ellas:

Variables Geográficas:

- **COD.DEP:** Código único del departamento. Sirve para identificar la ubicación geográfica a nivel departamental.
- **DEPARTAMENTO:** Nombre del departamento en el cual se encuentra la información sobre el cultivo.
- **COD.MUN:** Código único del municipio dentro del departamento.

Variables de Cultivo:

- **MUNICIPIO:** Nombre del municipio en el cual se mide el rendimiento de los cultivos.

- **GRUPO DE CULTIVO:** Clasificación general del cultivo (por ejemplo, cereales, leguminosas, tubérculos, etc.).
- **SUBGRUPO DE CULTIVO:** Subcategoría dentro del grupo de cultivo, permitiendo un análisis más específico (por ejemplo, arroz dentro de cereales).
- **CULTIVO:** Nombre del cultivo específico, como maíz, café, caña de azúcar, etc.
- **DESAGRADACION REGIONAL Y/O SISTEMA PRODUCTIVO:** Información relacionada con la región geográfica y el sistema productivo utilizado en el área de cultivo.

Variables de Producción:

- **AREA SEMBRADA (ha):** Superficie total sembrada con el cultivo, medida en hectáreas.
- **AREA COSECHA (ha):** Superficie total cosechada de un cultivo, también medida en hectáreas.
- **PRODUCCION (t):** La cantidad total producida del cultivo, medida en toneladas.
- **RENDIMIENTO (t/ha):** El rendimiento del cultivo, expresado como la cantidad de producción por hectárea.
- **ESTADO FISICO PRODUCCION:** Describe el estado físico de la producción, lo que puede incluir detalles sobre la calidad o el estado de la cosecha.
- **NOMBRE CIENTIFICO:** El nombre científico del cultivo, usado en estudios más especializados y técnicos.
- **CICLO DE CULTIVO:** Información sobre el ciclo de vida del cultivo, que puede incluir detalles sobre la duración y las etapas de crecimiento.

Variables Temporales y Descriptivas:

- **AÑO:** El año en que se registró el rendimiento del cultivo, lo que permite evaluar las variaciones a lo largo del tiempo.
- **PERIODO:** Representa el período del año en que se midió el rendimiento (por ejemplo, si es temporada de siembra o cosecha).

Relevancia y Usos

Esta base de datos es de gran importancia para diversos actores del sector agrícola en Colombia. **Investigadores** pueden utilizar estos datos para estudiar las tendencias de productividad y los factores que afectan el rendimiento agrícola a lo largo del tiempo. **Políticos y tomadores de decisiones** pueden emplear esta información para desarrollar políticas agrícolas basadas en evidencia, que fomenten el uso eficiente de la tierra y los recursos en el país. Además, los **productores agrícolas** pueden usar estos datos para tomar decisiones informadas sobre qué cultivos sembrar, cuándo sembrarlos y cómo optimizar su rendimiento.

La base de datos también es útil para estudios relacionados con los **cambios climáticos** y cómo estos afectan los cultivos a lo largo de los años, así como para investigaciones sobre la eficiencia de los sistemas productivos y las prácticas agrícolas en las distintas regiones de Colombia.

b) Selección de Variable Principal y Análisis Descriptivo

Selección de la Variable Principal

Para este análisis, se ha seleccionado la variable **RENDIMIENTO (t/ha)** como la variable de mayor importancia en la base de datos. Esta selección se fundamenta en las siguientes razones:

- **Indicador clave de eficiencia:** El rendimiento por hectárea es la medida más importante para evaluar la eficiencia productiva de los cultivos.
- **Comparabilidad:** Permite realizar comparaciones objetivas entre diferentes cultivos, regiones y períodos de tiempo.
- **Relevancia económica:** Es fundamental para la toma de decisiones de inversión y políticas agrícolas.
- **Indicador de sostenibilidad:** Refleja la capacidad de producción de la tierra y su uso eficiente.

Objetivo del Análisis

Objetivo General: Caracterizar el comportamiento del rendimiento de cultivos (t/ha) en Colombia durante el período 2007-2018, identificando patrones, tendencias y variabilidad en la productividad agrícola a nivel nacional.

Objetivos Específicos: 1. Describir la distribución estadística del rendimiento de cultivos en Colombia. 2. Identificar los cultivos con mayor y menor rendimiento promedio. 3. Analizar la evolución temporal del rendimiento durante el período de estudio. 4. Examinar las diferencias regionales en el rendimiento agrícola.

Análisis Descriptivo del Rendimiento

Table 2: Estadísticas Descriptivas del Rendimiento (t/ha)

N	Media	Mediana	Desv.Est.	Mínimo	Máximo	Q1	Q3	CV
202635	9.24	5	14.89	0.03	246	1.5	11.23	161.15

Según los estadísticos de la Tabla 2, el rendimiento promedio de los cultivos en Colombia es de 9.24 t/ha, con una mediana de 5 t/ha. La diferencia entre media y mediana, junto con un coeficiente de variación elevado (161.15%), indica una fuerte dispersión y asimetría en los datos.

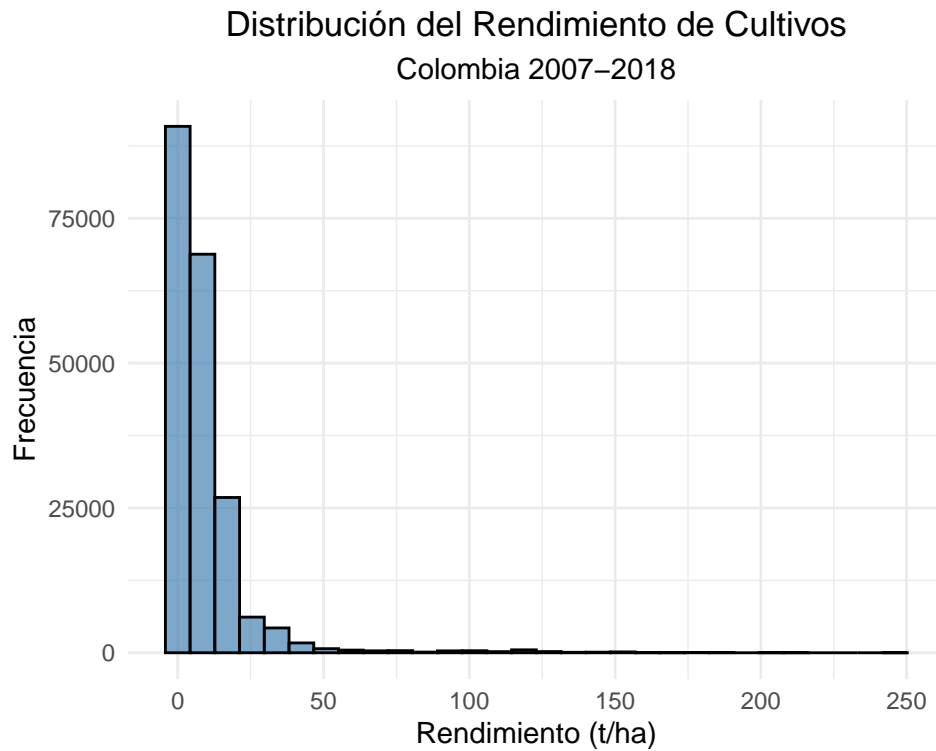
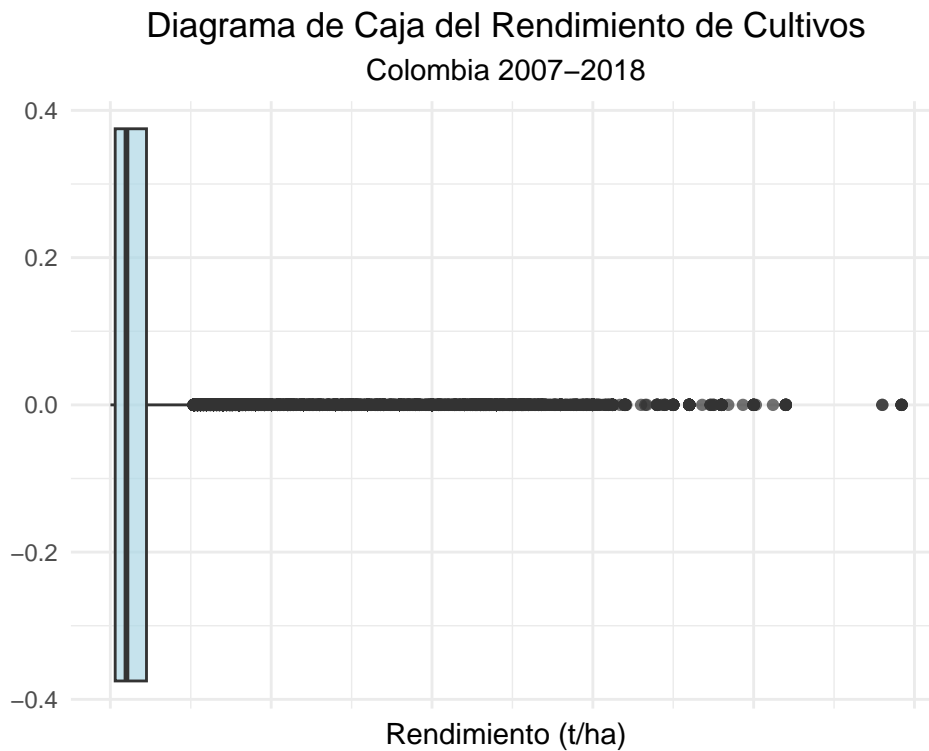


Figure 1: Histograma Rendimiento

La Figura 1(histograma) evidencia que la distribución del rendimiento está sesgada positivamente. La mayoría de los datos se concentran en valores bajos, mientras que existen pocos valores muy altos que funcionan como outliers.



La Figura ?? refuerza esta observación mostrando una gran cantidad de datos atípicos por encima del rango intercuartílico, confirmando la presencia de rendimientos excepcionalmente altos en ciertos cultivos.

Table 3: Top 10 Cultivos con Mayor Rendimiento Promedio

CULTIVO	Rendimiento_Promedio	N_observaciones
CAÑA AZUCARERA	117.30	619
TOMATE	38.29	9641
CHAMPIÑON	34.18	11
PIÑA	28.95	2125
SABILA	28.37	238
MAIZ FORRAJERO	28.23	234
CHAMBA	27.54	24
CEBOLLIN	25.64	14
FRESA	24.44	691
PAPAYA	23.29	1752

En cuanto a los cultivos más productivos, la Tabla 3 y su gráfica asociada muestran que la **caña de azúcar** lidera con un rendimiento promedio de 117.30 t/ha, seguida por cultivos como **tomate**, **champiñón**, **piña** y **sábila**, todos con rendimientos notablemente superiores al promedio general.

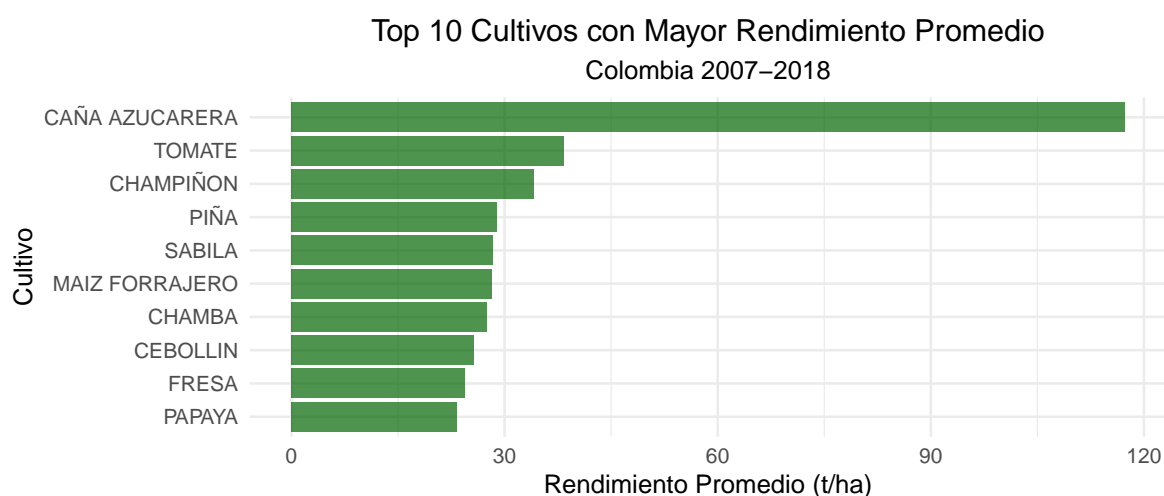


Figure 2: Histograma Rendimiento

Estos resultados de la Figura 2 permiten identificar cultivos estratégicos que, por su alto rendimiento, podrían ser priorizados en políticas de productividad agrícola o inversión técnica en regiones donde sean más comunes.