Análisis de los Grupos de Investigación más Reconocidos en Colombia: Un Enfoque Exploratorio con Machine Learning

Autores: Juan Pablo Alzate Villada, Romel Bayer
Institución Académica: TalentoTech CIAF
Curso: Inteligencia Artificial – Nivel exploratorio-Virtual

Docente: Directivos del curso – TalentoTech -CIAF

Fecha de entrega: 15 de Agosto del 2025

Historia y Contexto de la Inteligencia Artificial

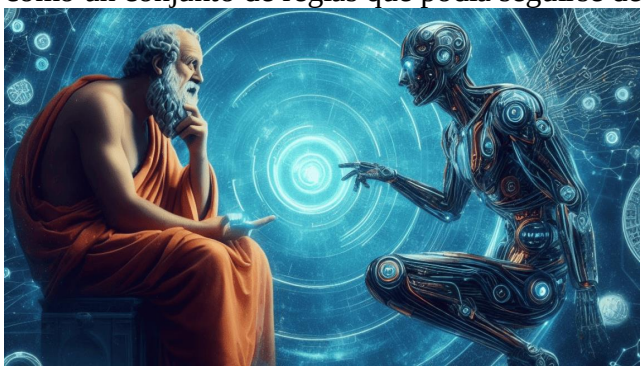
Introducción

La **Inteligencia Artificial (IA)** es una de las áreas más fascinantes y revolucionarias de la tecnología moderna. Se refiere a la capacidad de las máquinas para realizar tareas que, hasta hace poco, solo podían ser ejecutadas por seres humanos, tales como reconocer patrones, tomar decisiones, aprender de la experiencia y resolver problemas complejos. Para comprender el presente y futuro de la IA, es necesario recorrer su historia, entender sus avances, y reflexionar sobre los conceptos que la sustentan.

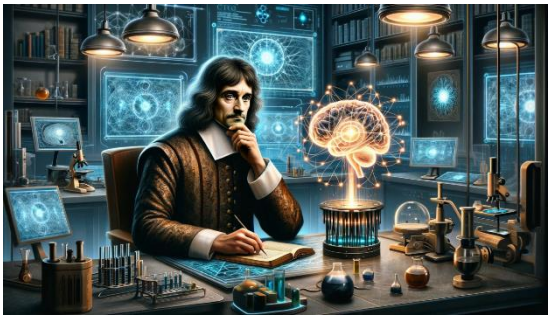
Breve historia y evolución

La idea de que las máquinas puedan pensar o razonar como humanos no es nueva; tiene raíces que se remontan a siglos atrás.

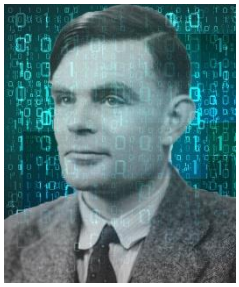
Antigüedad: Filósofos como Aristóteles ya exploraban el razonamiento lógico como un conjunto de reglas que podía seguirse de forma mecánica.



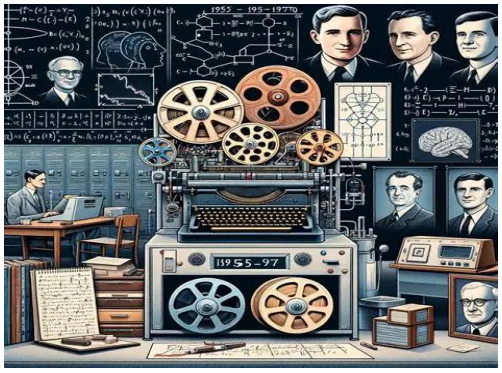
Siglo XVII: Matemáticos como René Descartes y Gottfried Leibniz imaginaron "máquinas de razonar" basadas en símbolos y lógica.



Década de 1940: Con la aparición de los primeros computadores electrónicos, científicos como Alan Turing comenzaron a preguntarse si estas máquinas podrían "pensar".

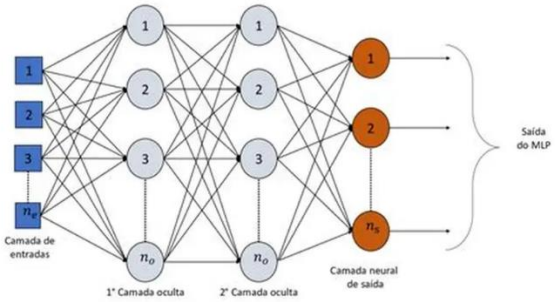


1956: Nace oficialmente el término *Artificial Intelligence* durante la conferencia de Dartmouth, donde John McCarthy, Marvin Minsky, Allen Newell y Herbert Simon sentaron las bases del campo.



1960-1970: Primeras aplicaciones como programas de ajedrez, sistemas expertos y chatbots primitivos (por ejemplo, *ELIZA* en 1966).

1980-1990: Auge de los **sistemas expertos** y primeras redes neuronales, aunque limitadas por la potencia de hardware de la época.



2000-2010: Avances gracias al *Machine Learning*, grandes volúmenes de datos (Big Data) y hardware más rápido. Aparecen aplicaciones en traducción automática, reconocimiento facial y asistentes virtuales.

2010-Actualidad: Revolución del *Deep Learning* con redes neuronales profundas capaces de realizar tareas como generar imágenes, escribir textos, diagnosticar enfermedades y conducir vehículos autónomos. Herramientas como ChatGPT, Midjourney o AlphaGo muestran el potencial de la IA en campos creativos y científicos

Tipos de inteligencia (según Howard Gardner)

El psicólogo Howard Gardner propuso la teoría de las **Inteligencias Múltiples**, donde cada persona posee varios tipos de inteligencia, en distinta proporción:

1. **Lógico-matemática** – Razonamiento lógico y resolución de problemas numéricos.
2. **Lingüística** – Uso del lenguaje oral y escrito.
3. **Espacial** – Visualización y orientación en el espacio.
4. **Musical** – Sensibilidad al ritmo, melodía y armonía.
5. **Corporal-kinestésica** – Control del cuerpo y habilidades físicas.
6. **Interpersonal** – Comprender y relacionarse con otras personas.
7. **Intrapersonal** – Conocimiento de uno mismo, autoconciencia.
8. **Naturalista** – Observación y clasificación del mundo natural.

En IA, estos tipos de inteligencia inspiran modelos que tratan de replicar ciertas capacidades humanas.

Tipos de inteligencia artificial

La IA se clasifica, según su capacidad, en tres niveles:

1. **IA Débil o Estrecha (ANI)** – Diseñada para tareas específicas (ej.: asistentes virtuales, traductores automáticos).
2. **IA General (AGI)** – Capaz de razonar, aprender y adaptarse a cualquier tarea intelectual como lo haría un humano (aún en desarrollo).
3. **IA Superinteligente (ASI)** – Hipotética IA que superaría ampliamente la inteligencia humana en todos los aspectos.

Tipos de aprendizaje en inteligencia artificial

La IA aprende a través de diferentes métodos:

1. **Aprendizaje supervisado** – Se entrena con datos etiquetados (ej.: clasificar correos como “spam” o “no spam”).
2. **Aprendizaje no supervisado** – Encuentra patrones en datos sin etiquetas (ej.: agrupar clientes por comportamiento).
3. **Aprendizaje por refuerzo** – Aprende mediante prueba y error, recibiendo recompensas o penalizaciones (ej.: robots aprendiendo a caminar).
4. **Aprendizaje semi-supervisado** – Combina datos etiquetados y no etiquetados para entrenar modelos.
5. **Aprendizaje profundo (Deep Learning)** – Usa redes neuronales con muchas capas para extraer representaciones complejas de los datos.

Conclusión

La Inteligencia Artificial ha pasado de ser una idea filosófica para convertirse en una herramienta indispensable en la vida moderna. Sus avances han transformado la medicina, la educación, el transporte, la comunicación y prácticamente todos los ámbitos humanos. El desafío para el futuro será equilibrar sus beneficios con el uso ético y responsable, asegurando que la IA se convierta en un aliado para el progreso y no en una amenaza.

Elaboración del Proyecto:

En este presente Documento explicaremos paso a paso la estructura del proyecto con base a los conocimientos dados en las clases virtuales correspondientes, aquí se detallara de manera exacta el desarrollo del machine leaning y los procesos de aprendisajez con la herramientas virtuales necesarias que vimos en el prensete documento, este proyecto se trata de analizar, visualizar , y usar todos esos dataframes para generarlos en los entorno virtuales correspondientes, donde utilizamos como fuente principal el lenguaje Python como bibliotecas de ultima generación de vanguardia donde nos permite usar esas herramientas junto al entorno (Google colab) donde nos facilitara ya que las bibliotecas están integradas de manera por defecto lo que nos ahorra la instalaciones de dichas, en las clases virtuales hemos adquirido un buen conjunto de herramientas como:

Python :

Python es una de mis lenguajes favoritos ya que es ideal para el aprendizaje automatizado, Python en el momento es uno de los lenguajes de programación mas utilizados , y Python es ideal para el manejo de datos , Python te permite utilizar una gran gama de herramientas que se basan en este lenguaje como es pandas, matploit, numpy etc..

Para Descargar Python:

Welcome to Python.org | Download Python | Python.org

Pandas:

Analiza datos y es mas flexible al generar tablas o informacion correspondientes de dataframes que con llevan a un tipo de generación de visualización o función necesaria que van a correponder en dicha función, pandas es muy recursivo lo que permite trabajar con poca líneas de código, lo que es muy factible ya que no consumiría y no obstruiría visualimente el análisis y la carga de datos , pandas trabaja de manera conjunta y unida con otras bibliotecas como matplotlib , entre otras, permite cargar archivos csv, xlsx , doc , sql , etc..

Importación

```
import pandas as pd # Juan Pablo Alzate Vilada:  Importacion de pandas
```

en Google colab no es necesaria la instalación, ya que Google colab trae por defecto pandas integrado simplemente usamos “ import pandas as pd” ya a breviamos como “pd” Pan – Das que es una contracción de Panel-Data

DataFrame:

```
df = pd.read csv('/content/Grupos de Investigaci n Reconocidos 20250814.csv')
```

En una variable llamada df que es nuestro dataframe asignamos las funcionalidades de pandas

estamos diciendo que la variable df con la función pandas lea (read) archivos csv (y adentro esta el archivo csv correspondiente)

-Inspeccion Inicial:

Inspección inicial de los datos

Antes de realizar cualquier análisis, se llevó a cabo una **inspección inicial** del archivo Grupos_de_Investigaci_n_Reconocidos_20250814.csv con el fin de conocer la estructura del conjunto de datos, identificar los tipos de variables, detectar posibles valores faltantes y obtener una visión general de la información disponible.

Para ello, se utilizaron los métodos info() y describe() de la librería **Pandas**:

- **df.info()** permitió obtener información general, incluyendo:

  Número total de filas y columnas.
  Nombres de las columnas.
  Tipo de dato de cada columna (numérico, texto, fecha, etc.).
  Cantidad de valores no nulos por columna.

Uso de memoria del DataFrame.

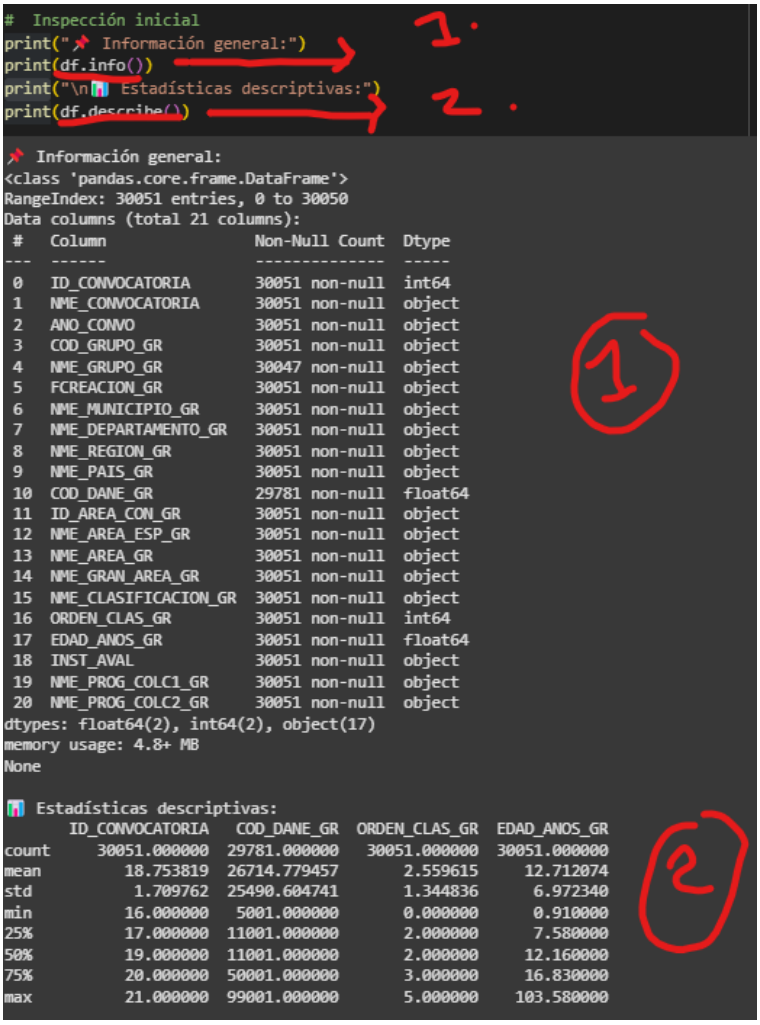
- **df.describe()** generó estadísticas descriptivas para las columnas numéricas:

  Count: cantidad de valores no nulos.

  Mean: promedio.

  Std: desviación estándar (medida de dispersión).

  Min, 25%, 50%, 75%, Max: valores mínimo, cuartiles y valor máximo.



Esta revisión inicial permitió:

1. Confirmar que el archivo fue cargado correctamente.
2. Identificar columnas que contienen valores faltantes.
3. Conocer el rango y la distribución de los valores numéricos.
4. Definir posibles pasos de limpieza y análisis posteriores.

!El dataset contiene **30.051 registros** y **21 columnas**, de las cuales **17 son de tipo texto (object)**, **2 son numéricas enteras (int64)** y **2 son numéricas decimales (float64)**. La memoria utilizada es de aproximadamente **4,8 MB!**

ID_CONVOCATORIA: Oscila entre 16 y 21, con un promedio de 18,75, lo que podría indicar diferentes ediciones o convocatorias a lo largo del tiempo.

COD_DANE_GR: Códigos geográficos que van desde 5001 hasta 99001, con valores repetidos en percentiles intermedios, lo que sugiere concentración en ciertas regiones.

ORDEN_CLAS_GR: Varía de 0 a 5, con un promedio cercano a 2,56, posiblemente indicando niveles de clasificación de los grupos.

EDAD_ANOS_GR: La edad de los grupos oscila entre 0,91 y 103,58 años, con un promedio de 12,71 años, lo que indica que la mayoría son relativamente recientes, pero existen casos con larga trayectoria.

-FILTRADO PRIMERAS FILAS (50)

```
[ ]  df.head(50)
```

df.head(50)

	ID_CONVOCATORIA	NME_CONVOCATORIA	ANO_CONVO	COD_GRUPO_GR	NME_GRUPO_GR	FCREACION_GR	NME_MUNICIPIO_GR	NME_DEPARTAMENTO_GR	NME_REGION_GR	NME_PAIS_GR	...	ID_AREA_CON_GR	NME_AREA_ESP_GR	NME_AREA_GR	NME_GRAN_AREA_GR	NME_CLASIFICACION_GR	ORDEN_CLAS_GR	EDAD_ANOS_GR	INST_AVAL	NME_PROG_COLC1_GR	NME_PROG_COLC2_GR
0	16	Convocatoria 640 de 2013	31/10/2013	COL0016283	Socialización y violencia	01/02/1993	Bogotá, D.C.	Bogotá, D. C.	Distrito Capital	Colombia	...	5I	No registra	Otras Ciencias Sociales	Ciencias Sociales	C	2	21.00	UNIVERSIDAD CENTRAL	Ciencia, Tecnología e Innovación en Ciencias H...	Ciencia, Tecnología e Innovación en Educación
1	16	Convocatoria 640 de 2013	31/10/2013	COL0018751	ECITRONICA	01/01/2002	Bogotá, D.C.	Bogotá, D. C.	Distrito Capital	Colombia	...	2B	No registra	Ingenierías Eléctrica, Electrónica e Informática	Ingeniería y Tecnología	C	2	12.08	ESCUELA COLOMBIANA DE INGENIERIA JULIO GARAVITO	Ciencia, Tecnología e Innovación en Tecnología...	Desarrollo Tecnológico e Innovación Industrial
2	16	Convocatoria 640 de 2013	31/10/2013	COL0013334	Línea de Investigación en Jóvenes y Culturas J...	01/09/1995	Bogotá, D.C.	Bogotá, D. C.	Distrito Capital	Colombia	...	5D	No registra	Sociología	Ciencias Sociales	D	1	18.67	UNIVERSIDAD CENTRAL	Ciencia, Tecnología e Innovación en Ciencias H...	Ciencia, Tecnología e Innovación en Educación
3	16	Convocatoria 640 de 2013	31/10/2013	COL0013316	Género y Cultura	01/09/1998	Bogotá, D.C.	Bogotá, D. C.	Distrito Capital	Colombia	...	5I	No registra	Otras Ciencias Sociales	Ciencias Sociales	C	2	15.42	UNIVERSIDAD CENTRAL	Ciencia, Tecnología e Innovación en Ciencias H...	Ciencia, Tecnología e Innovación en Educación
4	16	Convocatoria 640 de 2013	31/10/2013	COL0014583	Pavimentos y Materiales de Ingeniería	01/12/2002	Bogotá, D.C.	Bogotá, D. C.	Distrito Capital	Colombia	...	2A	No registra	Ingeniería Civil	Ingeniería y Tecnología	D	1	11.17	UNIVERSIDAD CATOLICA DE COLOMBIA	Ciencia, Tecnología e innovación en Ambiente, ...	Desarrollo Tecnológico e Innovación Industrial
5	16	Convocatoria 640 de 2013	31/10/2013	COL0008764	Grupo de Investigacion en Gestión Industrial GEGI	01/09/2002	Bogotá, D.C.	Bogotá, D. C.	Distrito Capital	Colombia	...	2K	No registra	Otras Ingenierías y Tecnologías	Ingeniería y Tecnología	D	1	11.67	UNIVERSIDAD CATOLICA DE COLOMBIA	Desarrollo Tecnológico e Innovación Industrial	Ciencia, Tecnología e Innovación en Ciencias H...
6	16	Convocatoria 640 de 2013	31/10/2013	COL0020571	Palabra, pueblo y vida	01/12/1998	Bogotá, D.C.	Bogotá, D. C.	Distrito Capital	Colombia	...	6E	No registra	Otras Humanidades	Humanidades	D	1	15.17	CORPORACION UNIVERSITARIA MINUTO DE DIOS UNIMI...	Ciencia, Tecnología e Innovación en Ciencias H...	No Aplica
7	16	Convocatoria 640 de 2013	31/10/2013	COL0005584	GIP	01/01/1995	Bogotá, D.C.	Bogotá, D. C.	Distrito Capital	Colombia	...	2K	No registra	Otras Ingenierías y Tecnologías	Ingeniería y Tecnología	C	2	19.08	UNIVERSIDAD CATOLICA DE COLOMBIA	Desarrollo Tecnológico e Innovación Industrial	No Aplica
8	16	Convocatoria 640 de 2013	31/10/2013	COL0022164	Grupo de Virología	01/01/1999	Bogotá, D.C.	Bogotá, D. C.	Distrito Capital	Colombia	...	3B	No registra	Medicina Clínica	Ciencias Médicas y de la Salud	A1	5	15.08	UNIVERSIDAD EL BOSQUE	Ciencia, Tecnología e Innovación en Salud	Biotecnología
9	16	Convocatoria 640 de 2013	31/10/2013	COL0024534	Investigación y Desarrollo en Operación del Tr...	01/09/2003	Tunja	Boyacá	Centro Oriente	Colombia	...	2A	No registra	Ingeniería Civil	Ingeniería y Tecnología	B	3	10.67	UNIVERSIDAD PEDAGOGICA Y TECNOLOGICA DE COLOMBIA	Ciencia, Tecnología e Innovación en Ciencias H...	Ciencia, Tecnología e innovación en Ambiente, ...
10	16	Convocatoria 640 de 2013	31/10/2013	COL0024122	Lógica, Epistemología y Filosofía de la Ciencia	01/08/2002	Bogotá, D.C.	Bogotá, D. C.	Distrito Capital	Colombia	...	6E	No registra	Otras Humanidades	Humanidades	A1	5	11.50	UNIVERSIDAD DE LOS ANDES | COLEGIO MAYOR NUEST...	Ciencia, Tecnología e Innovación en Ciencias H...	No Aplica
11	16	Convocatoria 640 de 2013	31/10/2013	COL0021292	Grupo Investigación Economía Social (GRIES)	01/03/2002	Bogotá, D.C.	Bogotá, D. C.	Distrito Capital	Colombia	...	5B	No registra	Economía y Negocios	Ciencias Sociales	Reconocido	0	11.92	UNIVERSIDAD SANTO TOMAS	Ciencia, Tecnología e Innovación en Ciencias H...	No Aplica
12	16	Convocatoria 640 de 2013	31/10/2013	COL0014313	PEDAGOGIA Y LENGUAJES	01/02/2000	Bucaramanga	Santander	Centro Oriente	Colombia	...	5C	No registra	Ciencias de la Educación	Ciencias Sociales	D	1	14.00	UNIVERSIDAD COOPERATIVA DE COLOMBIA	Ciencia, Tecnología e Innovación en Ciencias H...	Ciencia, Tecnología e Innovación en Educación
13	16	Convocatoria 640 de 2013	31/10/2013	COL0021372	GRUBIODEQ(Grupo de Biotecnología. Depto de Qui...	01/09/2003	Montería	Córdoba	Caribe	Colombia	...	1D	No registra	Ciencias Químicas	Ciencias Naturales	B	3	10.67	UNIVERSIDAD DE CORDOBA	Biotecnología	Ciencias Básicas
14	16	Convocatoria 640 de 2013	31/10/2013	COL0016209	Sistemas Complejos	01/05/2000	Bogotá, D.C.	Bogotá, D. C.	Distrito Capital	Colombia	...	1C	No registra	Ciencias Físicas	Ciencias Naturales	A1	5	13.75	UNIVERSIDAD ANTONIO NARIÑO	Ciencias Básicas	Ciencia, Tecnología e Innovación en Salud
15	16	Convocatoria 640 de 2013	31/10/2013	COL0023509	Gestión de la construcción	01/02/2001	Medellín	Antioquia	Eje Cafetero	Colombia	...	2A	No registra	Ingeniería Civil	Ingeniería y Tecnología	D	1	13.00	UNIVERSIDAD EAFIT	Desarrollo Tecnológico e Innovación Industrial	Ciencia, Tecnología e innovación en Ambiente, ...
16	16	Convocatoria 640 de 2013	31/10/2013	COL0006788	Geofísica	01/01/2001	Bogotá, D.C.	Bogotá, D. C.	Distrito Capital	Colombia	...	1E	No registra	Ciencias de la Tierra y Medioambientales	Ciencias Naturales	C	2	13.08	UNIVERSIDAD ANTONIO NARIÑO	Ciencias Básicas	Ciencia, Tecnología e innovación en Ambiente, ...
17	16	Convocatoria 640 de 2013	31/10/2013	COL0024294	Sociedad, ciencia y tecnología en Colombia	01/07/2000	Bogotá, D.C.	Bogotá, D. C.	Distrito Capital	Colombia	...	5B	No registra	Economía y Negocios	Ciencias Sociales	B	3	13.58	OBSERVATORIO COLOMBIANO DE CIENCIA Y TECNOLOGÍ...	Ciencia, Tecnología e Innovación en Ciencias H...	No Aplica
18	16	Convocatoria 640 de 2013	31/10/2013	COL0001137	Biotecnología UNIAGRARIA	01/12/2000	Bogotá, D.C.	Bogotá, D. C.	Distrito Capital	Colombia	...	2D	No registra	Ingeniería Química	Ingeniería y Tecnología	C	2	13.17	FUNDACION UNIVERSITARIA AGRARIA DE COLOMBIA	Biotecnología	Desarrollo Tecnológico e Innovación Industrial
19	16	Convocatoria 640 de 2013	31/10/2013	COL0022511	Grupo de Biofísica	01/02/1992	Medellín	Antioquia	Eje Cafetero	Colombia	...	1F	No registra	Ciencias Biológicas	Ciencias Naturales	C	2	22.00	UNIVERSIDAD NACIONAL DE COLOMBIA	Ciencias Básicas	No Aplica
20	16	Convocatoria 640 de 2013	31/10/2013	COL0015919	SIDRe - Sistemas de Información, Sistemas Dist...	01/01/2000	Bogotá, D.C.	Bogotá, D. C.	Distrito Capital	Colombia	...	2B	No registra	Ingenierías Eléctrica, Electrónica e Informática	Ingeniería y Tecnología	B	3	14.08	PONTIFICIA UNIVERSIDAD JAVERIANA	Ciencia, Tecnología e Innovación en Tecnología...	Desarrollo Tecnológico e Innovación Industrial
21	16	Convocatoria 640 de 2013	31/10/2013	COL0020974	ASOCIACION: CENTRO DE ESTUDIOS REGIONALES, REGION	01/07/1993	Cali	Valle del Cauca	Pacífico	Colombia	...	6A	No registra	Historia y Arqueología	Humanidades	A	4	20.58	UNIVERSIDAD DEL VALLE | UNIVERSIDAD PEDAGOGICA...	Ciencia, Tecnología e Innovación en Ciencias H...	No Aplica
22	16	Convocatoria 640 de 2013	31/10/2013	COL0023804	Emergencias y Desastres	01/02/1998	Medellín	Antioquia	Eje Cafetero	Colombia	...	3C	No registra	Ciencias de la Salud	Ciencias Médicas y de la Salud	C	2	18.00	UNIVERSIDAD DE ANTIOQUIA	Ciencia, Tecnología e Innovación en Ciencias H...	Ciencia, Tecnología e Innovación en Salud
23	16	Convocatoria 640 de 2013	31/10/2013	COL0000908	Biofísica y bioquímica estructural	01/11/1999	Bogotá, D.C.	Bogotá, D. C.	Distrito Capital	Colombia	...	1C	No registra	Ciencias Físicas	Ciencias Naturales	C	2	14.25	PONTIFICIA UNIVERSIDAD JAVERIANA	Ciencias Básicas	Ciencia, Tecnología e Innovación en Salud
24	16	Convocatoria 640 de 2013	31/10/2013	COL0010833	Grupo de Peliculas delgadas y Nanofotónica P.U.J.	01/02/1998	Bogotá, D.C.	Bogotá, D. C.	Distrito Capital	Colombia	...	1C	No registra	Ciencias Físicas	Ciencias Naturales	B	3	18.00	PONTIFICIA UNIVERSIDAD JAVERIANA	Ciencias Básicas	Ciencia, Tecnología e Innovación en Tecnología...
25	16	Convocatoria 640 de 2013	31/10/2013	COL0009135	Grupo de Investigación Fitoquímica Universidad...	01/05/1976	Bogotá, D.C.	Bogotá, D. C.	Distrito Capital	Colombia	...	1D	No registra	Ciencias Químicas	Ciencias Naturales	B	3	37.67	PONTIFICIA UNIVERSIDAD JAVERIANA	Ciencias Básicas	Ciencia, Tecnología e Innovación en Ciencias A...
26	16	Convocatoria 640 de 2013	31/10/2013	COL0011839	Ciencia e Ingeniería del agua y el ambiente	01/01/1999	Bogotá, D.C.	Bogotá, D. C.	Distrito Capital	Colombia	...	1E	No registra	Ciencias de la Tierra y Medioambientales	Ciencias Naturales	A1	5	15.08	PONTIFICIA UNIVERSIDAD	Ciencia, Tecnología e innovación en Ambiente,	Ciencias Básicas

27	16	Convocatoria 640 de 2013	31/10/2013	COL0022889	Grupo de Física Teórica y Computacional	01/01/2002	Tunja	Boyacá	Centro Oriente	Colombia	...	1C	No registra	Ciencias Físicas	Ciencias Naturales	B	3	12.08	UNIVERSIDAD PEDAGOGICA Y TECNOLOGICA DE COLOMBIA
28	16	Convocatoria 640 de 2013	31/10/2013	COL0006759	Genética de Poblaciones Molecular y Biología E...	01/01/1997	Bogotá, D.C.	Bogotá, D. C.	Distrito Capital	Colombia	...	1F	No registra	Ciencias Biológicas	Ciencias Naturales	C	2	17.08	PONTIFICIA UNIVERSIDAD JAVERIANA
29	16	Convocatoria 640 de 2013	31/10/2013	COL0023484	GISA	01/01/1999	Riohacha	La Guajira	Caribe	Colombia	...	2G	No registra	Ingeniería Ambiental	Ingeniería y Tecnología	B	3	15.08	UNIVERSIDAD DE LA GUAJIRA
30	16	Convocatoria 640 de 2013	31/10/2013	COL0014009	TERAPIA CELULAR Y MOLECULAR	01/01/2000	Bogotá, D.C.	Bogotá, D. C.	Distrito Capital	Colombia	...	1F	No registra	Ciencias Biológicas	Ciencias Naturales	B	3	14.08	PONTIFICIA UNIVERSIDAD JAVERIANA
31	16	Convocatoria 640 de 2013	31/10/2013	COL0004489	Estructuras y construcción	01/01/2002	Bogotá, D.C.	Bogotá, D. C.	Distrito Capital	Colombia	...	2A	No registra	Ingeniería Civil	Ingeniería y Tecnología	A1	5	12.08	PONTIFICIA UNIVERSIDAD JAVERIANA
32	16	Convocatoria 640 de 2013	31/10/2013	COL0000935	Bioingeniería, análisis de señales y procesami...	01/01/1998	Bogotá, D.C.	Bogotá, D. C.	Distrito Capital	Colombia	...	2B	No registra	Ingenierías Eléctrica, Electrónica e Informática	Ingeniería y Tecnología	B	3	16.08	PONTIFICIA UNIVERSIDAD JAVERIANA
33	16	Convocatoria 640 de 2013	31/10/2013	COL0015885	SIRP - Sistemas Inteligentes, Robótica y Perce...	01/01/2000	Bogotá, D.C.	Bogotá, D. C.	Distrito Capital	Colombia	...	2B	No registra	Ingenierías Eléctrica, Electrónica e Informática	Ingeniería y Tecnología	C	2	14.08	PONTIFICIA UNIVERSIDAD JAVERIANA
34	16	Convocatoria 640 de 2013	31/10/2013	COL0001244	CECATA	01/08/2000	Bogotá, D.C.	Bogotá, D. C.	Distrito Capital	Colombia	...	2A	No registra	Ingeniería Civil	Ingeniería y Tecnología	A	4	13.50	PONTIFICIA UNIVERSIDAD JAVERIANA
35	16	Convocatoria 640 de 2013	31/10/2013	COL0018878	Grupo de Investigaciones Contables y Gestión P...	01/07/2002	Medellín	Antioquia	Eje Cafetero	Colombia	...	5B	No registra	Economía y Negocios	Ciencias Sociales	C	2	11.58	UNIVERSIDAD DE MEDELLIN
36	16	Convocatoria 640 de 2013	31/10/2013	COL0014939	Procesos sociales y salud	01/02/2000	Bogotá, D.C.	Bogotá, D. C.	Distrito Capital	Colombia	...	3C	No registra	Ciencias de la Salud	Ciencias Médicas y de la Salud	B	3	14.00	PONTIFICIA UNIVERSIDAD JAVERIANA
37	16	Convocatoria 640 de 2013	31/10/2013	COL0002733	Conceptualizacion y Práctica de Enfermeria	01/01/2000	Bogotá, D.C.	Bogotá, D. C.	Distrito Capital	Colombia	...	3C	No registra	Ciencias de la Salud	Ciencias Médicas y de la Salud	C	2	14.08	PONTIFICIA UNIVERSIDAD JAVERIANA
38	16	Convocatoria 640 de 2013	31/10/2013	COL0021908	FISQUIM - Fisicoquímica de Geofluídos	01/01/1996	Cali	Valle del Cauca	Pacífico	Colombia	...	1E	No registra	Ciencias de la Tierra y Medioambientales	Ciencias Naturales	C	2	18.08	SERVICIO GEOLOGICO COLOMBIANO | UNIVERSIDAD DE...
39	16	Convocatoria 640 de 2013	31/10/2013	COL0005913	Grupo de Instrumentación Científica & Didáctica	01/10/1995	Bogotá, D.C.	Bogotá, D. C.	Distrito Capital	Colombia	...	1C	No registra	Ciencias Físicas	Ciencias Naturales	D	1	18.33	UNIVERSIDAD DISTRITAL FRANCISCO JOSE DE CALDAS
40	16	Convocatoria 640 de 2013	31/10/2013	COL0010207	Grupo de Investigación en Psicología y Salud	01/07/2001	Bogotá, D.C.	Bogotá, D. C.	Distrito Capital	Colombia	...	3C	No registra	Ciencias de la Salud	Ciencias Médicas y de la Salud	A1	5	12.58	PONTIFICIA UNIVERSIDAD JAVERIANA
41	16	Convocatoria 640 de 2013	31/10/2013	COL0016877	UNIDAD DE INVESTIGACIONES AGROPECUARIAS	01/01/2002	Bogotá, D.C.	Bogotá, D. C.	Distrito Capital	Colombia	...	4A	No registra	Agricultura, Silvicultura y Pesca	Ciencias Agrícolas	A1	5	12.08	PONTIFICIA UNIVERSIDAD JAVERIANA	Ciencia, Tecnología e Innovación en Ciencias A...	Ciencias Básicas
42	16	Convocatoria 640 de 2013	31/10/2013	COL0006929	CINNCO - Conocimiento, innovación y competitiv...	01/07/1995	Bogotá, D.C.	Bogotá, D. C.	Distrito Capital	Colombia	...	5B	No registra	Economía y Negocios	Ciencias Sociales	B	3	18.58	PONTIFICIA UNIVERSIDAD JAVERIANA	Ciencia, Tecnología e Innovación en Ciencias H...	Desarrollo Tecnológico e Innovación Industrial
43	16	Convocatoria 640 de 2013	31/10/2013	COL0014725	Política Económica	01/07/2000	Bogotá, D.C.	Bogotá, D. C.	Distrito Capital	Colombia	...	5B	No registra	Economía y Negocios	Ciencias Sociales	B	3	13.58	PONTIFICIA UNIVERSIDAD JAVERIANA	Ciencia, Tecnología e Innovación en Ciencias H...	Desarrollo Tecnológico e Innovación Industrial
44	16	Convocatoria 640 de 2013	31/10/2013	COL0022001	KINEPAIDEIA	01/01/1998	Bucaramanga	Santander	Centro Oriente	Colombia	...	5C	No registra	Ciencias de la Educación	Ciencias Sociales	C	2	16.08	UNIVERSIDAD COOPERATIVA DE COLOMBIA	Ciencia, Tecnología e Innovación en Ciencias H...	No Aplica
45	16	Convocatoria 640 de 2013	31/10/2013	COL0004774	Estética, nuevas tecnologías y habitabilidad	01/07/1998	Bogotá, D.C.	Bogotá, D. C.	Distrito Capital	Colombia	...	6D	No registra	Arte	Humanidades	A	4	15.58	PONTIFICIA UNIVERSIDAD JAVERIANA	Ciencia, Tecnología e innovación en Ambiente, ...	Ciencia, Tecnología e Innovación en Ciencias H...
46	16	Convocatoria 640 de 2013	31/10/2013	COL0005708	Grupo de Investigaciones en Materiales y Estru...	01/04/2002	Bogotá, D.C.	Bogotá, D. C.	Distrito Capital	Colombia	...	6D	No registra	Arte	Humanidades	C	2	11.83	PONTIFICIA UNIVERSIDAD JAVERIANA	Ciencias Básicas	Ciencia, Tecnología e Innovación en Ciencias H...
47	16	Convocatoria 640 de 2013	31/10/2013	COL0000812	Integración y contexto Contable	01/05/1995	Bogotá, D.C.	Bogotá, D. C.	Distrito Capital	Colombia	...	5B	No registra	Economía y Negocios	Ciencias Sociales	C	2	18.75	PONTIFICIA UNIVERSIDAD JAVERIANA	Ciencia, Tecnología e Innovación en Ciencias H...	Desarrollo Tecnológico e Innovación Industrial
48	16	Convocatoria 640 de 2013	31/10/2013	COL0021089	Ecología de Bosques Andinos Colombianos - EBAC	01/08/2001	Tunja	Boyacá	Centro Oriente	Colombia	...	1F	No registra	Ciencias Biológicas	Ciencias Naturales	C	2	12.50	UNIVERSIDAD PEDAGOGICA Y TECNOLOGICA DE COLOMBIA	Ciencias Básicas	Ciencia, Tecnología e innovación en Ambiente, ...
49	16	Convocatoria 640 de 2013	31/10/2013	COL0024463	Inmunobiología y Biología Celular	01/01/1996	Bogotá, D.C.	Bogotá, D. C.	Distrito Capital	Colombia	...	1F	No registra	Ciencias Biológicas	Ciencias Naturales	B	3	18.08	PONTIFICIA UNIVERSIDAD JAVERIANA	Ciencias Básicas	Ciencia, Tecnología e Innovación en Salud

50 rows × 21 columns

Análisis inicial con df.head(50)

Se utilizó el método head(50) de la librería *Pandas* para visualizar las primeras 50 filas del dataset. Este análisis permitió:

1. **Revisar la estructura general de la base de datos**:
observar cómo están organizadas las columnas y qué tipo de información contiene cada una (fechas, nombres, códigos, clasificaciones, etc.).
2. **Identificar los campos principales**:
por ejemplo, datos de identificación de los grupos de investigación, su ubicación geográfica, área de conocimiento, clasificación y antigüedad.
3. **Detectar posibles inconsistencias o errores iniciales**:
formatos de fecha diferentes, valores faltantes, códigos incompletos o duplicados.

4. **Familiarizarse con el contenido:**

   antes de aplicar técnicas de limpieza y análisis estadístico.

Gracias a esta inspección, se pudo confirmar que la base de datos contiene tanto variables cualitativas (nombres, áreas de investigación, clasificación) como cuantitativas (edad del grupo, año de convocatoria, orden de clasificación), lo que permitirá aplicar análisis mixtos en las etapas posteriores.

Valores Faltantes:

```
df.isnull().sum()
```

Análisis de valores nulos (`df.isnull().sum()`)

Se aplicó el método `isnull().sum()` para contar cuántos valores faltantes (`NaN`) hay en cada columna del dataset. Este análisis permitió:

1. **Identificar columnas con datos incompletos**, lo que es clave para decidir estrategias de limpieza (relleno, eliminación de filas, etc.).
2. **Detectar la columna más afectada**: por ejemplo, en este dataset la variable `COD_DANE_GR` presenta valores nulos, mientras que otras como `NME_GRUPO_GR` tienen ausencias menores.
3. **Determinar la calidad de los datos**: al observar que la mayoría de las columnas no presentan nulos, se puede concluir que la base de datos tiene buena completitud general, pero con ciertas excepciones que deben corregirse antes del análisis estadístico o modelado.

Análisis de valores nulos (`df.isnull().sum()`)

Se realizó un conteo de valores faltantes en cada columna de la base de datos. Los resultados muestran que:

	0
ID_CONVOCATORIA	0
NME_CONVOCATORIA	0
ANO_CONVO	0
COD_GRUPO_GR	0
NME_GRUPO_GR	4
FCREACION_GR	0
NME_MUNICIPIO_GR	0
NME_DEPARTAMENTO_GR	0
NME_REGION_GR	0
NME_PAIS_GR	0
COD_DANE_GR	270
ID_AREA_CON_GR	0
NME_AREA_ESP_GR	0
NME_AREA_GR	0
NME_GRAN_AREA_GR	0
NME_CLASIFICACION_GR	0
ORDEN_CLAS_GR	0
EDAD_ANOS_GR	0
INST_AVAL	0
NME_PROG_COLC1_GR	0
NME_PROG_COLC2_GR	0

dtype: int64

- La mayoría de las columnas no presentan datos nulos, lo que indica una **alta completitud** de la información.
- La columna **NME_GRUPO_GR** presenta **4 valores nulos**, lo que representa una ausencia mínima dentro del dataset.
- La columna **COD_DANE_GR** es la más afectada, con **270 valores nulos**, posiblemente debido a que algunos grupos no tienen registrado el código DANE del municipio correspondiente.

Este diagnóstico es importante porque:

1. **Confirma la calidad general del dataset**, ya que más del 90% de las columnas están completas.
2. **Permite planificar el tratamiento de valores faltantes**, decidiendo si se imputarán, se eliminarán filas o si se dejarán como están según su relevancia para el análisis.
3. **Evita problemas en cálculos y modelos posteriores**, ya que los valores nulos pueden afectar promedios, conteos y predicciones.

En resumen, el dataset está mayormente completo, pero requiere atención especial en **COD_DANE_GR** y **NME_GRUPO_GR** antes de realizar análisis avanzados.

Limpieza de Datos:

Limpieza de datos y eliminación de valores nulos

Con el objetivo de garantizar la calidad y consistencia del análisis, se procedió a eliminar todas las filas que contenían valores nulos utilizando el método dropna() de *Pandas*.

Pasos realizados:

1. **Identificación de valores nulos**:
previamente, con df.isnull().sum() se determinó que algunas columnas, como NME_GRUPO_GR y COD_DANE_GR, contenían datos faltantes.

2. **Aplicación de limpieza**:

```
df_limpio = df.dropna()
```

```python
# limpieza de datos

df_limpio = df.dropna()

print(df_limpio.isnull().sum())

df_limpio.to_csv("/content/Grupos_de_Investigacion_Limpio.csv", index=False)
print("\n Archivo limpio guardado como 'Grupos_de_Investigacion_Limpio.csv'")
```

Este comando elimina **todas las filas** que tengan al menos un valor faltante en cualquier columna.

3. **Verificación posterior**: se utilizó nuevamente df_limpio.isnull().sum() para confirmar que el nuevo DataFrame no contiene valores nulos.
4. **Exportación del dataset limpio**: el archivo resultante se guardó como Grupos_de_Investigacion_Limpio.csv para su uso en análisis posteriores.

Resultado:
Se obtuvo un dataset sin valores nulos, garantizando que los cálculos estadísticos, visualizaciones y modelos de aprendizaje automático no se vean afectados por datos incompletos.

```
ID_CONVOCATORIA         0
NME_CONVOCATORIA        0
ANO_CONVO               0
COD_GRUPO_GR            0
NME_GRUPO_GR            0
FCREACION_GR            0
NME_MUNICIPIO_GR        0
NME_DEPARTAMENTO_GR     0
NME_REGION_GR           0
NME_PAIS_GR             0
COD_DANE_GR             0
ID_AREA_CON_GR          0
NME_AREA_ESP_GR         0
NME_AREA_GR             0
NME_GRAN_AREA_GR        0
NME_CLASIFICACION_GR    0
ORDEN_CLAS_GR           0
EDAD_ANOS_GR            0
INST_AVAL               0
NME_PROG_COLC1_GR       0
NME_PROG_COLC2_GR       0
dtype: int64

 Archivo limpio guardado como 'Grupos_de_Investigacion_Limpio.csv'
```

Verificación de limpieza de datos

Luego de aplicar el procedimiento de eliminación de valores nulos (dropna()), se realizó una nueva verificación con el comando:

df_limpio.isnull().sum()

El resultado mostró que **todas las columnas tienen 0 valores nulos**, lo que confirma que el conjunto de datos está completamente limpio y listo para análisis.

Además, se exportó el archivo limpio con el nombre **Grupos_de_Investigacion_Limpio.csv**, lo que permite mantener una copia depurada para futuros usos.

Conclusión de esta etapa:

- Los datos están íntegros y sin valores faltantes.
- Se garantiza que los análisis estadísticos y modelos posteriores no se verán afectados por registros incompletos.

Normalización y estandarización de variables numéricas

```python
# Variables Normalizadas

#  Seleccionar solo columnas numéricas
df_numericas = df.select_dtypes(include=["int64", "float64"])

# Normalización Min-Max (valores entre 0 y 1)
scaler_minmax = MinMaxScaler()
df_minmax = pd.DataFrame(scaler_minmax.fit_transform(df_numericas),
                         columns=df_numericas.columns)

# 4. Estandarización Z-score (media=0, desviación estándar=1)
scaler_std = StandardScaler()
df_std = pd.DataFrame(scaler_std.fit_transform(df_numericas),
                      columns=df_numericas.columns)

# 5. Mostrar ejemplos
print("Normalización Min-Max:")
print(df_minmax.head())

print("\nEstandarización Z-score:")
print(df_std.head())
```

Normalización y estandarización de variables numéricas

Para preparar los datos antes de realizar análisis estadísticos y algoritmos de machine learning, se aplicaron dos técnicas de escalado sobre las variables numéricas del conjunto de datos:

1. **Normalización Min-Max**

   Transforma los valores para que estén en un rango entre **0 y 1**.

Fórmula:

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Utilidad: evita que variables con rangos diferentes dominen los análisis.

2. **Estandarización Z-score**

$$Z = \frac{X - \mu}{\sigma}$$

- Ajusta los datos para que tengan **media = 0** y **desviación estándar = 1**.
- Fórmula:
-

  Utilidad: facilita el trabajo de algoritmos que suponen datos distribuidos normalmente.

Resultados

- La **normalización Min-Max** llevó todos los valores a una escala de 0 a 1, conservando la proporción entre ellos.
- La **estandarización Z-score** centró los datos en torno a cero, con valores positivos y negativos que representan desviaciones estándar respecto a la media.

Esto garantiza que las variables numéricas estén en escalas comparables, evitando sesgos en los análisis posteriores como **clustering, PCA o modelos predictivos**.

```
Normalización Min-Max:
   ID_CONVOCATORIA  COD_DANE_GR  ORDEN_CLAS_GR  EDAD_ANOS_GR
0              0.0      0.06383            0.4      0.195675
1              0.0      0.06383            0.4      0.108795
2              0.0      0.06383            0.2      0.172981
3              0.0      0.06383            0.4      0.141327
4              0.0      0.06383            0.2      0.099932

Estandarización Z-score:
   ID_CONVOCATORIA  COD_DANE_GR  ORDEN_CLAS_GR  EDAD_ANOS_GR
0        -1.610671    -0.616464      -0.416129      1.188706
1        -1.610671    -0.616464      -0.416129     -0.090656
2        -1.610671    -0.616464      -1.159726      0.854523
3        -1.610671    -0.616464      -0.416129      0.388388
4        -1.610671    -0.616464      -1.159726     -0.221174
```

◈ Normalización Min-Max

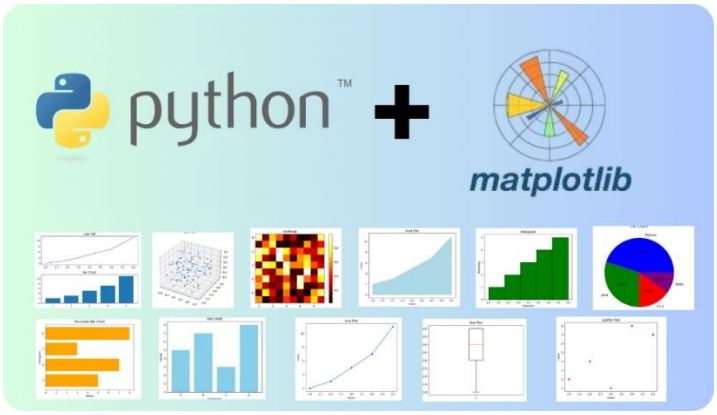
- Todos los valores quedaron entre **0 y 1**.

- Ejemplo: EDAD_ANOS_GR pasó de valores como *25, 40, etc.* a proporciones como 0.195675 o 0.108795.
- Útil para algoritmos que dependen de la distancia (K-Means, KNN, redes neuronales).

◈ Estandarización Z-score

- La media de cada columna es **0** y la desviación estándar es **1**.
- Los valores negativos indican que están por debajo de la media, y los positivos por encima.
- Ejemplo: EDAD_ANOS_GR tiene 1.188706 (1.18 desviaciones estándar sobre la media) y -0.090656 (ligeramente por debajo de la media).

-MATPLOTLIB:



Matplotlib es una biblioteca de Python para la creación de gráficos y visualizaciones de datos en 2D (y de forma limitada en 3D).
Permite generar desde gráficos simples (líneas, barras, pastel) hasta visualizaciones más personalizadas, con control total sobre colores, estilos, etiquetas y ejes.

Se usa mucho en análisis de datos y ciencia, y junto con **pandas** y **NumPy** forma parte del “ecosistema” principal para análisis y visualización en Python.

Grafica univariada:

value_counts() → Cuenta cuántas veces aparece cada gran área.

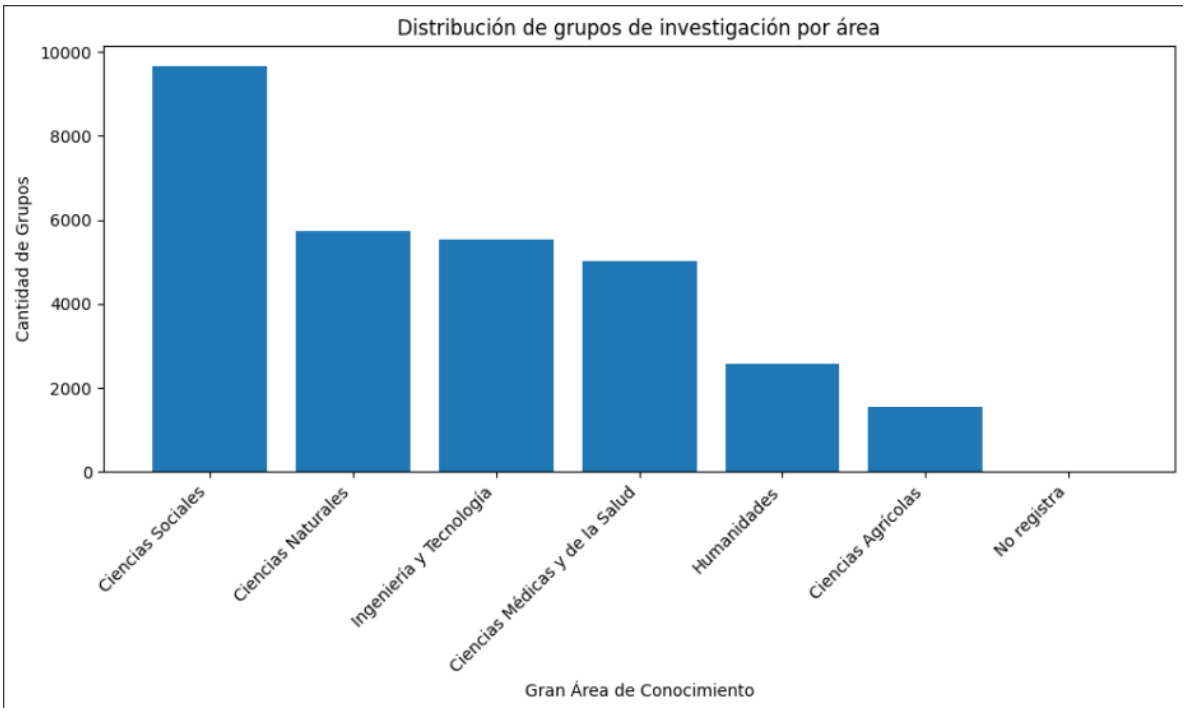
reset_index() → Convierte ese conteo en un DataFrame para graficarlo fácilmente.

plt.bar() → Dibuja las barras.

plt.xticks(rotation=45) → Gira las etiquetas para que no se encimen.

tight_layout() → Ajusta la gráfica para que todo se vea completo.

```
# 1 grafica Univariada

areas_investigacion = df["NME_GRAN_AREA_GR"].value_counts().reset_index()
areas_investigacion.columns = ["Gran Área de Conocimiento", "Cantidad_Grupos"]

plt.figure(figsize=(10,6))
plt.bar(areas_investigacion["Gran Área de Conocimiento"],areas_investigacion["Cantidad_Grupos"])
plt.xticks(rotation=45, ha='right')
plt.xlabel("Gran Área de Conocimiento")
plt.ylabel("Cantidad de Grupos")
plt.title("Distribución de grupos de investigación por área")
plt.tight_layout()
plt.show()
```



Análisis de la distribución de grupos de investigación por área

La gráfica muestra la cantidad de grupos de investigación distribuidos según la **Gran Área de Conocimiento**.
Se observan las siguientes tendencias:

Ciencias Sociales encabeza la lista con una cifra cercana a **9.700 grupos**, lo que evidencia que esta área concentra la mayor parte de los esfuerzos investigativos. Esto puede estar relacionado con la alta demanda de estudios en temas sociales, económicos, educativos y culturales.

Ciencias Naturales y **Ingeniería y Tecnología** ocupan el segundo y tercer lugar, con valores similares (alrededor de **5.700 y 5.500 grupos**, respectivamente). Esto sugiere un equilibrio entre investigación básica en ciencias y desarrollo tecnológico.

Ciencias Médicas y de la Salud cuenta con aproximadamente **5.000 grupos**, lo que refleja un interés considerable en áreas como medicina, salud pública y biotecnología aplicada a la salud.

Humanidades y **Ciencias Agrícolas** muestran cantidades menores (cerca de **2.600 y 1.500 grupos**, respectivamente), lo que indica que reciben menor atención comparativa, a pesar de su relevancia en contextos culturales y productivos.

La categoría **"No registra"** es prácticamente nula, lo que significa que la mayoría de los grupos están correctamente clasificados en alguna de las áreas.

Grafico de Densidad

```python
# Convertir la columna a numérica y eliminar no numéricos
df["EDAD_ANOS_GR"] = pd.to_numeric(df["EDAD_ANOS_GR"], errors="coerce")
edad_grupos = df["EDAD_ANOS_GR"].dropna()

# Gráfico de densidad
plt.figure(figsize=(8,5))
sns.kdeplot(x=edad_grupos, fill=True, color="skyblue")
plt.title("Densidad de la distribución de la antigüedad de los grupos de investigación")
plt.xlabel("Antigüedad (años)")
plt.ylabel("Densidad")
plt.grid(True, linestyle="--", alpha=0.5)
plt.show()

# Estadísticas descriptivas
edad_stats = {
    "mínimo": round(edad_grupos.min(), 2),
    "máximo": round(edad_grupos.max(), 2),
    "promedio": round(edad_grupos.mean(), 2),
    "mediana": round(edad_grupos.median(), 2),
    "desviación estándar": round(edad_grupos.std(), 2)
}

print("📊 Estadísticas de antigüedad de grupos:")
print(edad_stats)
```

limpieza de datos:

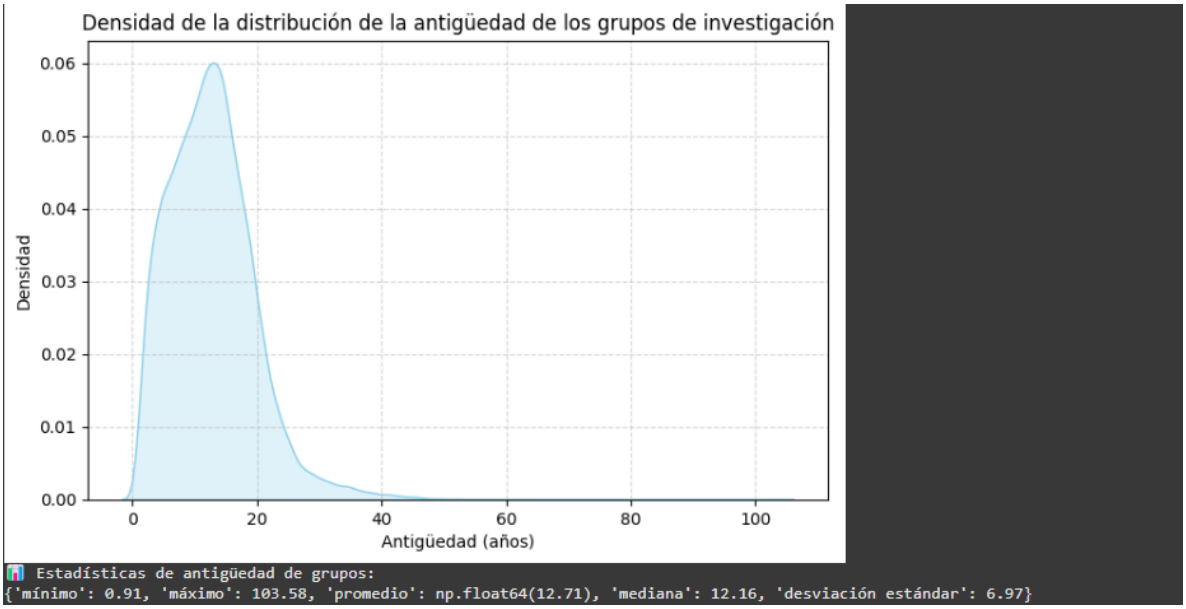
```python
# Convertir la columna a numérica y eliminar no numéricos
df["EDAD_ANOS_GR"] = pd.to_numeric(df["EDAD_ANOS_GR"], errors="coerce")
edad_grupos = df["EDAD_ANOS_GR"].dropna()
```

Grafico Densidad:

```python
# Gráfico de densidad
plt.figure(figsize=(8,5))
sns.kdeplot(x=edad_grupos, fill=True, color="skyblue")
plt.title("Densidad de la distribución de la antigüedad de los grupos de investigación")
plt.xlabel("Antigüedad (años)")
plt.ylabel("Densidad")
plt.grid(True, linestyle="--", alpha=0.5)
plt.show()
```

Calculo Estadistico:

```python
# Estadísticas descriptivas
edad_stats = {
    "mínimo": round(edad_grupos.min(), 2),
    "máximo": round(edad_grupos.max(), 2),
    "promedio": round(edad_grupos.mean(), 2),
    "mediana": round(edad_grupos.median(), 2),
    "desviación estándar": round(edad_grupos.std(), 2)
```



Interpretación visual

Interpretación visual

La **curva de densidad** muestra que la mayoría de los grupos de investigación tienen **antigüedad entre 5 y 20 años**.

Hay un **pico máximo** alrededor de los **10–15 años**, lo que indica que es la edad más común para un grupo activo.

La curva presenta una **cola larga hacia la derecha** (sesgo positivo), lo que significa que existen algunos grupos con antigüedades muy altas, incluso superiores a 40 y hasta más de 100 años, pero son casos muy poco frecuentes.

- Hay muy pocos grupos con menos de 3 años de existencia.

Conclusiones clave

1. **Concentración en antigüedad media** → La mayoría de los grupos tienen entre 10 y 15 años.
2. **Sesgo a la derecha** → Algunos grupos son muy antiguos, pero no representan la norma.
3. **Posible outlier** → El valor de **103 años** es inusual y podría revisarse para confirmar si es un error o un dato real.
4. **Estructura estable** → La distribución sugiere que hay una base sólida de grupos con trayectoria consolidada, lo que indica continuidad en la investigación.

Interpretación numérica

Métrica	Valor	Interpretación
Mínimo	0.91 años	Grupo más joven tiene menos de 1 año.
Máximo	103.58 años	Caso extremo, probablemente un grupo histórico o un error de registro.
Promedio	12.71 años	En promedio, los grupos tienen más de una década de trayectoria.
Mediana	12.16 años	La mitad de los grupos tiene menos de 12,16 años y la otra mitad más.
Desviación estándar	6.97 años	Hay una variabilidad moderada en las edades; no todos los grupos son de la misma generación.

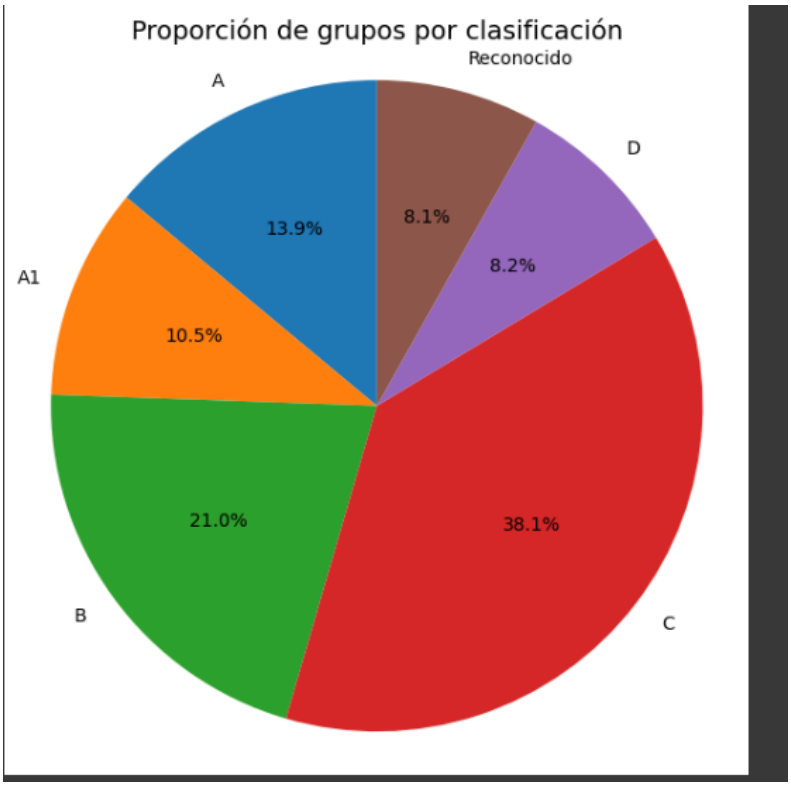
Grafico Pastel:

```python
# Contar grupos por clasificación

clasificacion_counts = df['NME_CLASIFICACION_GR'].value_counts().sort_index()

# Crear gráfico de pastel

plt.figure(figsize=(7,7))
plt.pie(
    clasificacion_counts,
    labels=clasificacion_counts.index,
    autopct='%1.1f%%',
    startangle=90
)
plt.title("Proporción de grupos por clasificación", fontsize=14)
plt.axis('equal')  # Mantener el pastel circular
plt.show()
```

Cada porción representa la **proporción de grupos** en una categoría de clasificación.

El uso de plt.axis('equal') asegura que el pastel no se deforme, manteniendo las proporciones reales.

La etiqueta autopct='%1.1f%%' añade el porcentaje con un decimal, lo que ayuda a interpretar rápidamente el peso de cada categoría.



Análisis:

La clasificación **C** es la más predominante, con **38.1%** del total, lo que indica que más de un tercio de los grupos pertenecen a este nivel.

La segunda clasificación más común es la **B** con **21.0%**, representando aproximadamente una quinta parte de los grupos.

Las clasificaciones **A** y **A1** tienen una participación menor, con **13.9%** y **10.5%** respectivamente.

Las clasificaciones **D** y **Reconocido** son las menos frecuentes, con porcentajes similares (**8.2%** y **8.1%**).

El patrón sugiere una **concentración en clasificaciones medias (C y B)**, mientras que los niveles más bajos y más altos están menos representados.

Grafico Bivariada

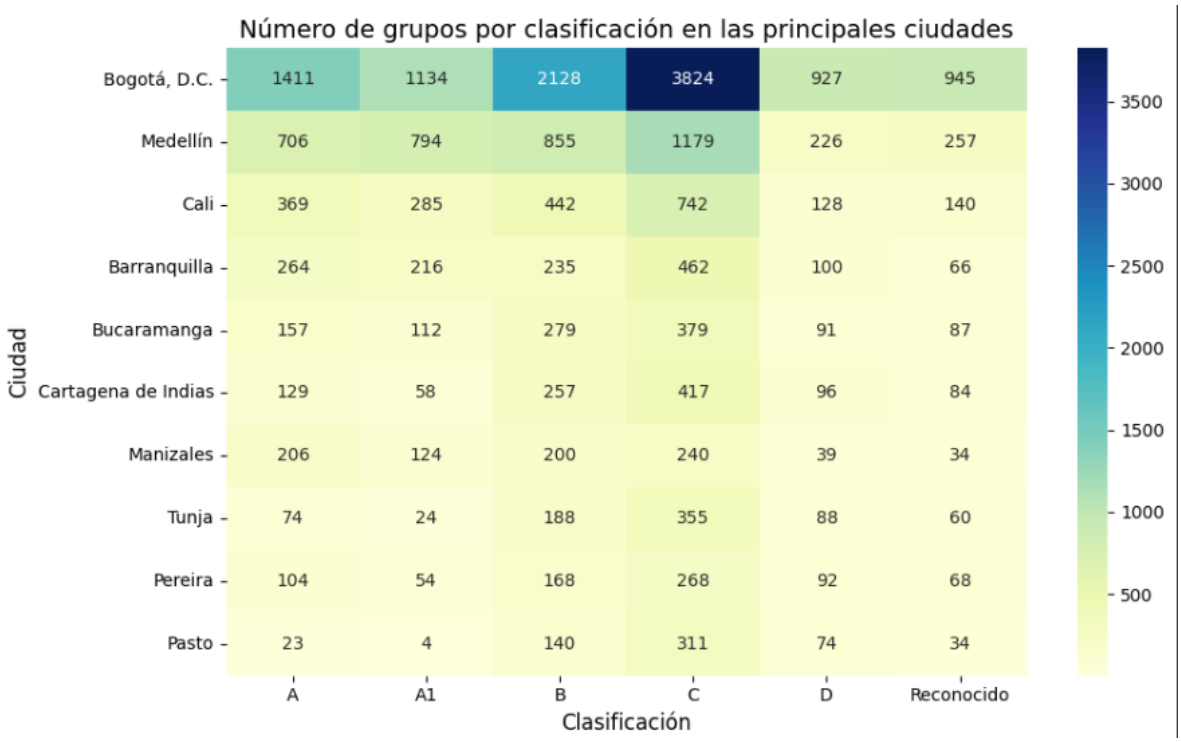
```python
# 1. grafica Bivariada

#  Creacion de tabla de ciudades y clasificaciones
clas_ciudad = pd.crosstab(df['NME_MUNICIPIO_GR'], df['NME_CLASIFICACION_GR'])

# Filtracion de las 10 ciudades con más grupos

top_ciudades = clas_ciudad.sum(axis=1).sort_values(ascending=False).head(10).index
clas_ciudad_top = clas_ciudad.loc[top_ciudades]

#  Generacion de mapa de calor
plt.figure(figsize=(10,6))
sns.heatmap(clas_ciudad_top, annot=True, fmt="d", cmap="YlGnBu")
plt.title("Número de grupos por clasificación en las principales ciudades", fontsize=14)
plt.xlabel("Clasificación", fontsize=12)
plt.ylabel("Ciudad", fontsize=12)
plt.tight_layout()
plt.show()
```

Análisis de la Gráfica Bivariada: Número de grupos por clasificación en las principales ciudades

La visualización presenta un mapa de calor que muestra la distribución de grupos de investigación según su clasificación en las diez ciudades con mayor cantidad de grupos. El color de las celdas indica la intensidad de la frecuencia: tonos más oscuros representan un mayor número de grupos en esa categoría.

Se observa que las clasificaciones **C** y **B** predominan en la mayoría de las ciudades, lo que sugiere una amplia base de grupos en niveles intermedios. Por otro lado, la presencia de grupos en categorías **A** y **A1** es más reducida y concentrada en ciertas ciudades, lo que indica un posible foco de excelencia en esos lugares.

Este análisis permite identificar patrones geográficos en la clasificación de los grupos, así como posibles áreas de oportunidad para fortalecer la calidad de la investigación en regiones donde las clasificaciones superiores son menos frecuentes.



Este mapa de calor muestra cómo se distribuye el número de **grupos por clasificación** en las 10 ciudades con mayor cantidad de registros.
Aquí tienes un análisis detallado:

1. Ciudad con mayor concentración

Bogotá D.C. lidera ampliamente en todas las clasificaciones, especialmente en **C** (3,824 grupos), seguida de **B** (2,128 grupos) y **A** (1,411 grupos).

Esto refleja que la capital concentra la mayor cantidad de grupos, probablemente por su tamaño poblacional y centralización de actividades.

2. Segundo grupo de ciudades importantes

Medellín y **Cali** son las que siguen en volumen.

Medellín tiene su punto más alto en la clasificación **C** (1,179 grupos).

Cali muestra un patrón similar, aunque con valores más bajos (742 grupos en C).

Ambas ciudades mantienen un número más equilibrado entre A, A1, B y C.

3. Ciudades intermedias

Barranquilla, Bucaramanga y Cartagena de Indias presentan valores intermedios.

Todas destacan más en la categoría **C**, pero con menos de 500 grupos.

Las clasificaciones **D** y **Reconocido** son bajas en comparación.

4. Ciudades con menor número de grupos

Manizales, Tunja, Pereira y Pasto tienen valores mucho menores.

La clasificación **C** sigue siendo la más frecuente, pero con cifras alrededor de los 300 o menos.

En algunas ciudades como **Pasto**, las categorías A y A1 casi no están presentes.

5. Patrones generales

La **clasificación C** es la más frecuente en casi todas las ciudades, indicando que probablemente sea el nivel más común de los grupos evaluados.

Las clasificaciones **A y A1** tienden a concentrarse más en ciudades grandes.

La categoría **Reconocido** y la **D** tienen menor representación en general.

6. Posibles interpretaciones

Las ciudades más grandes tienden a tener diversidad de clasificaciones y un número alto de grupos en todos los niveles.

El predominio de la clasificación **C** podría indicar que es un estándar intermedio donde se agrupa la mayor parte de las organizaciones o instituciones.

En ciudades pequeñas, hay poca variedad y cantidad de grupos, lo que podría estar relacionado con menor infraestructura o población.

Si quieres, puedo hacerte también un **análisis porcentual** para ver qué tan concentrada está cada clasificación dentro de cada ciudad, así se ve más claro el peso relativo.

treemap interactivo

```python
# --- Treemap interactivo ---
import plotly.express as px
fig = px.treemap(
    df,
    path=["CIUDAD", "NOMBRE_ACTOR"],  # Jerarquía: Ciudad -> Actor
    values=None,  # Todos del mismo tamaño
    color="CIUDAD",  # Colores según ciudad
    title="Treemap de actores reconocidos por ciudad",
    width=1000,
    height=700
)

# Mostrar gráfico
fig.show()
```

Jerarquía visual:

- El primer nivel son las **ciudades** (CIUDAD), representadas por bloques grandes.
- Dentro de cada bloque, hay subdivisiones por **nombre del actor** (NOMBRE_ACTOR).

Tamaño de los bloques:

- Como values=None, todos los actores dentro de una ciudad tienen el mismo tamaño, y el área total de la ciudad se reparte equitativamente.
- Esto significa que el tamaño refleja **cantidad de actores**, no una métrica numérica como presupuesto o población.

Colores:

- Se asigna un color distinto a cada ciudad (color="CIUDAD").
- Dentro de una misma ciudad, todos los actores tienen el mismo color, lo que refuerza la agrupación visual.

Interactividad:

- Puedes pasar el cursor por encima para ver **nombre de la ciudad, actor** y la posición jerárquica.
- Permite **hacer zoom** dentro de una ciudad y volver atrás.

Treemap de actores reconocidos por ciudad