**Logotipo

Descripción generada automáticamente**

**Inteligencia artificial aplicada a un foro de discusión tecnológica**

Julieth Tatiana García Zuluaga

Monografía presentada para optar al título de Especialista en Analítica y Ciencia de Datos

Asesor  
David Manuel Villanueva Valdes, Título académico más alto

Universidad de Antioquia  
Facultad de Ingeniería

Especialización en Analítica y Ciencia de Datos

Medellín, Antioquia, Colombia

2023

|  |  |
| --- | --- |
| **Cita** | (García Zuluaga, 2023) |
| **Referencia**  **Estilo APA 7 (2020)** | García Zuluaga, J. T. (2023). *Inteligencia Artificial aplicada a un foro de discusión tecnológica* Trabajo de grado especialización]. Universidad de Antioquia, Medellín, Colombia. |

**[](https://co.creativecommons.net/tipos-de-licencias/)** [](https://co.creativecommons.net/tipos-de-licencias/)

Especialización en Analítica y Ciencia de Datos, Cohorte VI.

Centro de Investigación Ambientales y de Ingeniería (CIA).

|  |  |
| --- | --- |
|  | Diagrama  Descripción generada automáticamente con confianza media |

Centro de Documentación Ingeniería (CENDOI)

**Repositorio Institucional:** http://bibliotecadigital.udea.edu.co

Universidad de Antioquia - www.udea.edu.co

Rector: John Jairo Arboleda Céspedes.

Decano: Julio Cesar Saldarriaga Molina

Jefe departamento: Diego José Luis Botia Valderrama

El contenido de esta obra corresponde al derecho de expresión de los autores y no compromete el pensamiento institucional de la Universidad de Antioquia ni desata su responsabilidad frente a terceros. Los autores asumen la responsabilidad por los derechos de autor y conexos.

**Tabla de contenido**

[Resumen 7](#_Toc151243029)

[Abstract 8](#_Toc151243030)

[1. Descripción del problema 9](#_Toc151243031)

[1.1. Problema de negocio 9](#_Toc151243032)

[1.2. Aproximación desde la analítica de datos 9](#_Toc151243033)

[1.3. Origen de los datos 10](#_Toc151243034)

[1.4. Métricas de desempeño 10](#_Toc151243035)

[2. Objetivos 12](#_Toc151243036)

[2.1. Objetivos generales 12](#_Toc151243037)

[2.2. Objetivos específicos 12](#_Toc151243038)

[3. Datos 13](#_Toc151243039)

[3.1. Datos originales 13](#_Toc151243040)

[3.2. Datsets 18](#_Toc151243041)

[Figura 1 Portada Normas APA séptima edición 2020 en inglés 19](#_Toc151243042)

[Referencias 21](#_Toc151243043)

**Lista de tablas**

[**Tabla 1**  Resultados del test PBQ-SF (Personality Belief Questionnaire Short Form) 19](#_Toc57037138)

[**Tabla 2** Características demográficas y tipo de tratamiento de hemodiálisis y diálisis peritoneal con la adherencia (SMAQ) 20](#_Toc57037139)

[**Tabla 3** Categorías de la investigación 21](#_Toc57037140)

**Lista de figuras**

[**Figura 1** Portada Normas APA séptima edición 2020 en inglés 22](#_Toc62551171)

[**Figura 2** Logo Universidad de Antioquia 22](#_Toc62551172)

**Siglas, acrónimos y abreviaturas**

**APA** American Psychological Association

**Cms.** Centímetros

**ERIC** Education Resources Information Center

**Esp.** Especialista

**MP** Magistrado Ponente

**MSc** Magister Scientiae

**Párr.** Párrafo

**PhD** Philosophiae Doctor

**PBQ-SF** Personality Belief Questionnaire Short Form

**PostDoc** PostDoctor

**UdeA** Universidad de Antioquia

# Resumen

La información presente en los foros de discusión pueden ser una fuente de información muy valiosa en las empresas para obtener un entendimiento más profundo de las necesidades y preferencias de los usuarios, identificando las fortalezas y debilidades de su conocimiento. Esto permitirá tomar decisiones más informadas, como la implementación de capacitaciones personalizadas, con el fin de mejorar las competencias necesarias y optimizar la gestión del conocimiento en la organización. Como resultado, se anticipa una mejora en la eficiencia y efectividad de los equipos, lo que se traducirá en un aumento de la productividad y la calidad de los proyectos desarrollados por la Vicepresidencia de Tecnología.

Este proyecto se enfoca en la aplicación de la inteligencia artificial para segmentar y clasificar a los usuarios que utilizan los foros de discusión en la Vicepresidencia de Tecnología de una empresa del sector financiero; así como la evaluación de la calidad de las respuestas proporcionadas por lo usuarios mediante la aplicación de un análisis de sentimientos y procesamiento de lenguaje natural.

Incluye al final de dicha página la dirección de los repositorios <https://github.com/TatianaGarcia1128/foro_discusion_udea> .

*Palabras clave*: Foros de discusión, Segmentación de usuarios, Análisis de sentimientos, Procesamiento de Lenguaje Natural.

# Abstract

The information present in discussion forums can be a valuable source of insights for companies to gain a deeper understanding of user needs and preferences, identifying strengths and weaknesses in their knowledge. This will enable more informed decision-making, such as implementing personalized training to enhance necessary skills and optimize knowledge management within the organization. Consequently, an improvement in team efficiency and effectiveness is anticipated, translating into increased productivity and the enhanced quality of projects developed by the Technology Vice Presidency.

This project focuses on the application of artificial intelligence to segment and classify users participating in discussion forums within the Technology Vice Presidency of a financial sector company. Additionally, it involves evaluating the quality of user responses through the application of sentiment analysis and natural language processing.

*Keywords***:** Discussion forums, User segmentation, Sentiment analysis, Natural Language Processing.

# Descripción del problema

En el contexto de la empresa, se implementó un foro de discusión hace más de un año con el propósito de abordar inquietudes y problemas cotidianos de los ingenieros de software. Hasta la fecha, no se ha aprovechado esta fuente de datos para respaldar la toma de decisiones estratégicas por parte de los líderes, particularmente en relación con la gestión del conocimiento y del talento en la organización, así como para maximizar el uso efectivo de la herramienta.

Estos datos ofrecen la posibilidad de obtener una comprensión profunda de las dificultades y desafíos específicos que enfrentan los equipos en el desarrollo de soluciones tecnológicas en diversas áreas de la Vicepresidencia de Tecnología. Al realizar un análisis detallado de estos datos, es factible identificar patrones, tendencias y áreas problemáticas específicas, proporcionando diagnósticos valiosos sobre el uso de la herramienta.

Para abordar esta problemática, se llevará a cabo una segmentación de usuarios mediante algoritmos de clusterización. Asimismo, se aplicarán técnicas de Procesamiento del Lenguaje Natural (PLN) y un análisis de sentimientos para evaluar la calidad de las respuestas dentro del foro de discusión y obtener un diagnóstico sobre el uso adecuado de la herramienta.

## Problema de negocio

Incorporar la inteligencia artificial en el análisis de los datos provenientes de un foro de discusión, con el fin de potenciar la toma de decisiones estratégicas centradas en la gestión del conocimiento, la administración del talento y la optimización de la herramienta.

## Aproximación desde la analítica de datos

Explica para qué servirían los modelos predictivos desarrollados. Por ejemplo: se desarrollarán modelos para recomendar productos a clientes.

Se aplicarán técnicas de Procesamiento del Lenguaje Natural (PLN) y un análisis de sentimientos para evaluar la calidad de las respuestas dentro del foro de discusión. De igual forma de aplicarán algoritmos de clusterización para la segmentación de usuarios que participan en el foro de discusión.

## Origen de los datos

El foro de discusión se ha desarrollado utilizando una herramienta basada en un proyecto de código abierto y emplea una base de datos Postgres para almacenar de manera persistente la información. Este foro se encuentra alojado en la nube de AWS, específicamente en el componente RDS Auto Scaling Group, junto con un Bucket S3.

Es importante destacar que el acceso a la base de datos no está́ disponible para el público en general. Para obtener copias de seguridad de la información almacenada o para llevar a cabo implementaciones específicas en la nube donde reside la base de datos, es necesario solicitar acceso a los administradores de bases de datos (DBAs). Esto puede incluir el uso de ETLs u otras estrategias para obtener la información necesaria.

La calidad de este foro está estrechamente relacionada con la naturaleza de las preguntas y respuestas presentadas. A menudo, se incorporan imágenes, lo que puede hacer que la extracción de información sea más laboriosa. Además, tanto las preguntas como las respuestas pueden contener fragmentos de código fuente y enlaces a otras fuentes de datos.

## Métricas de desempeño

Las métricas de machine learning que se usarán para la evaluación de la calidad de las respuestas son [1]:

* Polaridad (Positiva, Negativa, Neutral)
* Subjetividad
* Precisión y Exhaustividad
* Exactitud (Accuracy)
* F1-Score
* Matriz de confusión

Las métricas de machine learning que se usarán para la segmentación de usuarios son [2]:

* Calinski-Harabasz (CH)
* Índice de Rand Ajustado (ARI)
* Índice de información mutua (MI)
* Índice de Davies-Bouldin (DB)

La métrica de negocio relacionada con la evaluación de la calidad de las respuestas es: Tonalidad general de las respuestas proporcionadas en el foro de discusión.

Las métricas de negocio que se usarán para la segmentación de usuarios son: Nivel de participación y actividad de los usuarios en cada segmento.

A partir de un f1-score del 60% se estima que el modelo es lo suficientemente confiable para predecir la tonalidad general de las respuestas proporcionadas en el foro de discusión.

# Objetivos

## Objetivos generales

* Segmentar y clasificar a los usuarios que utilizan la herramienta de foro de discusión en la Vicepresidencia de Tecnología de la empresa.
* Evaluar la calidad de las respuestas proporcionadas por los usuarios en el foro de discusión en la Vicepresidencia de Tecnología de la empresa.

## Objetivos específicos

* Realizar procesamiento de texto para construir un dataset adecuado para la aplicación de los modelos de machine learning.
* Identificar los diferentes segmentos en los que se encuentran los usuarios que participan en el foro de discusión.
* Clasificar los segmentos anteriormente identificados de acuerdo a sus similitudes.
* Realizar un análisis de sentimientos para evaluar la calidad de las respuestas proporcionadas por los usuarios en el foro de discusión.

# Datos

## Datos originales

Los datos se recopilaron en un inicio haciendo una restauración local de un backup proporcionado por el DBA de la empresa que se encarga de administrar la base de datos del foro de discusión a mayo de 2023. Del total de 114 tablas disponibles en la base de datos restaurada se seleccionaron 12 tablas consideradas que contienen información relevante para la ejecución de los dos objetivos propuestos. Posterior a esta selección de tablas y campos se crea un query el cual es ejecutado mediante dos scripts en Python para exportar el resultado de las consultas en dos archivos .csv.

**Dataset para el objetivo de segmentación de usuarios**

El dataset que será usado presenta las siguientes características:

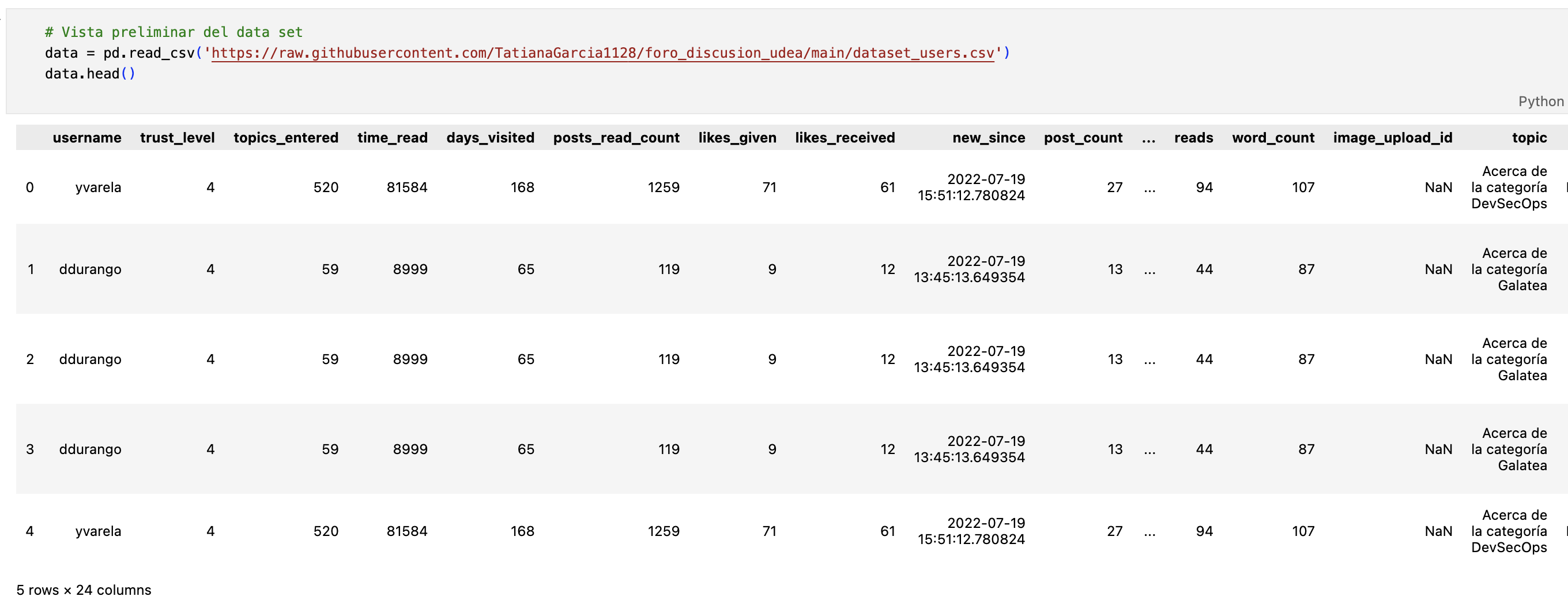


Imagen 1: Muestra dataset users (14 columnas de 24)

A continuación, se describe cada una de las columnas del dataset:

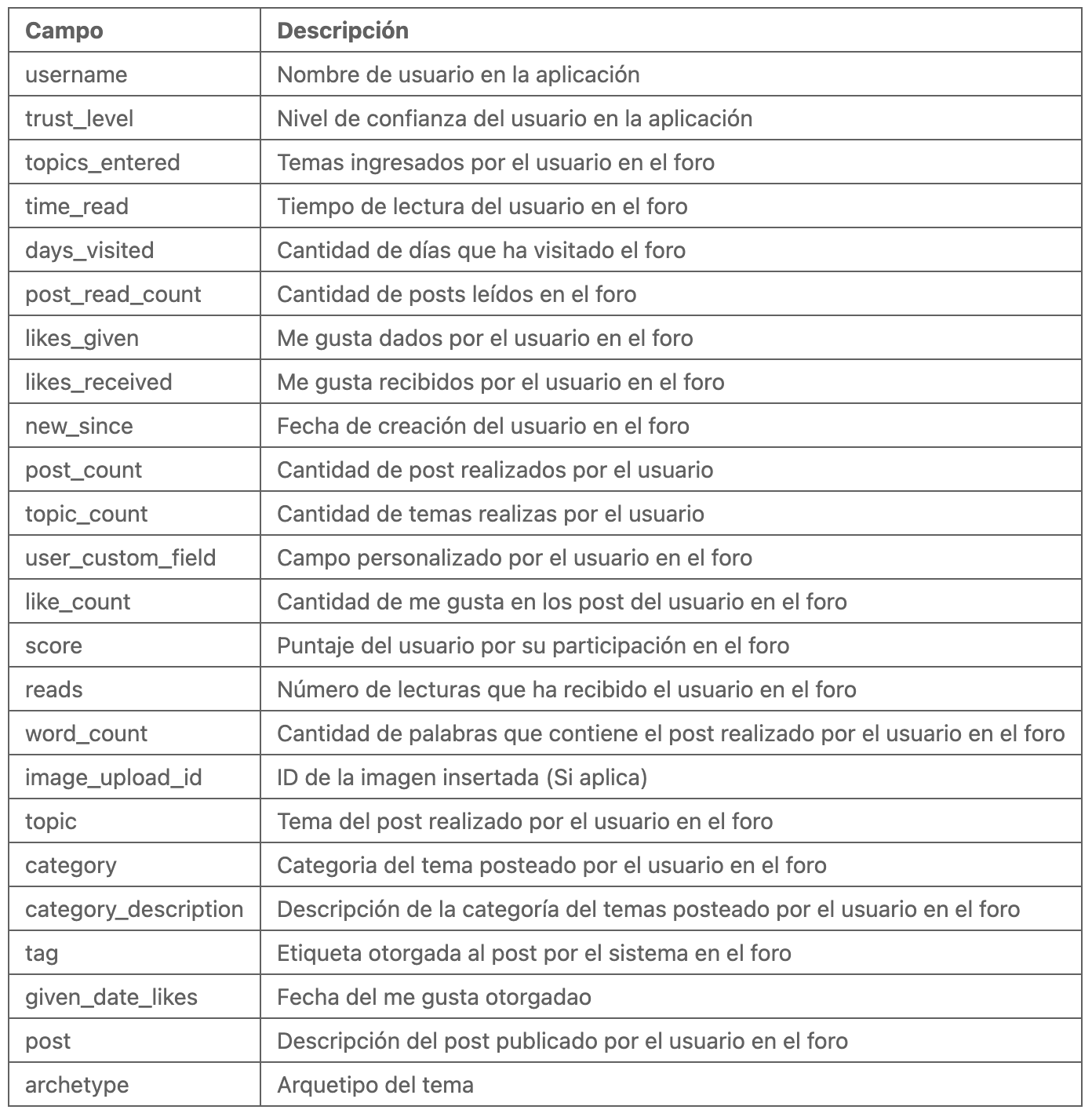


Imagen 2: Descripción de campos del dataset users

Este dataset tiene un total de 2.949 instancias y 24 columnas.

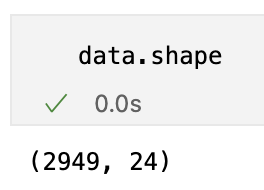
****

Imagen 3: Forma del dataset

El dataset contiene 2949 instancias y 24 columnas, de las cuales 12 tienen valores tipo entero (int64), 2 de tipo decimal (float64) y las 10 restantes son tipo object. Presenta una columna que en su totalidad esta nula, el campo ‘category\_description’ presenta valores nulos y en total el peso del dataset es de 553.1+ KB; es decir, que es fácil de procesar.

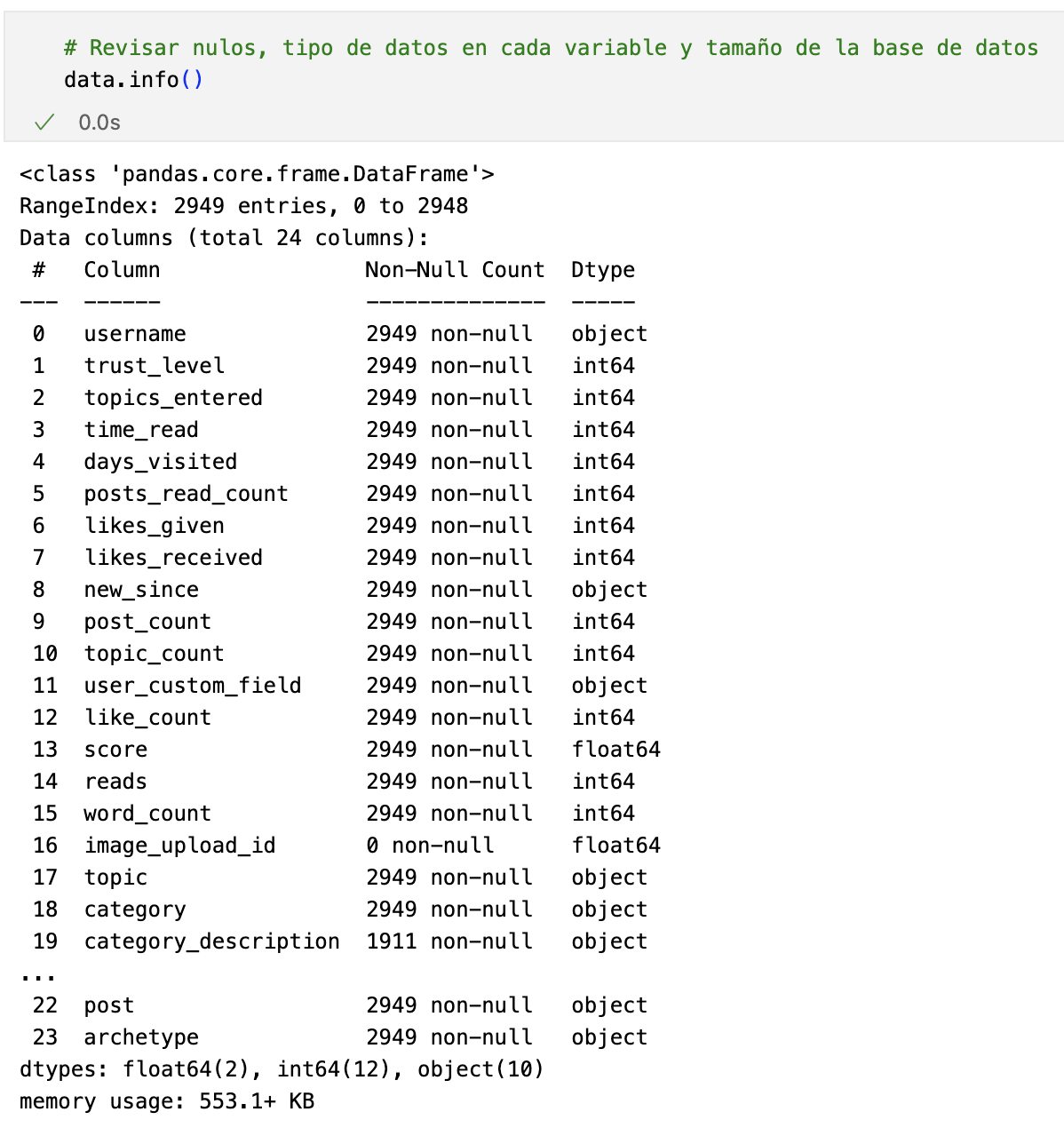


Imagen 4: Información detallada del dataset

El dataset presenta escalas diferentes por lo que posiblemente se requiera una estandarización de la información en pasos posteriores.

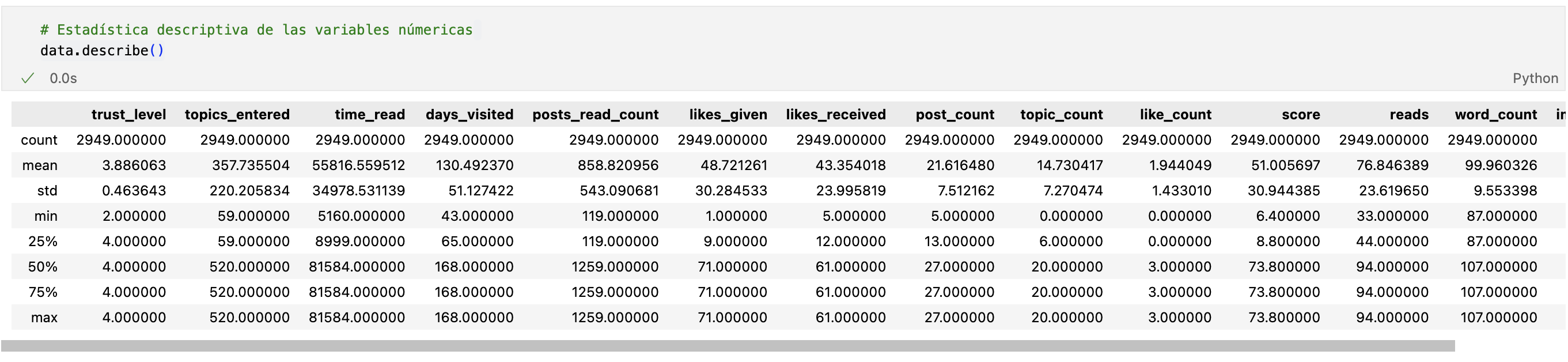


Imagen 5: Información descriptiva del dataset

**Dataset para la evaluación de la calidad de las respuestas**

El dataset presenta las siguientes características:

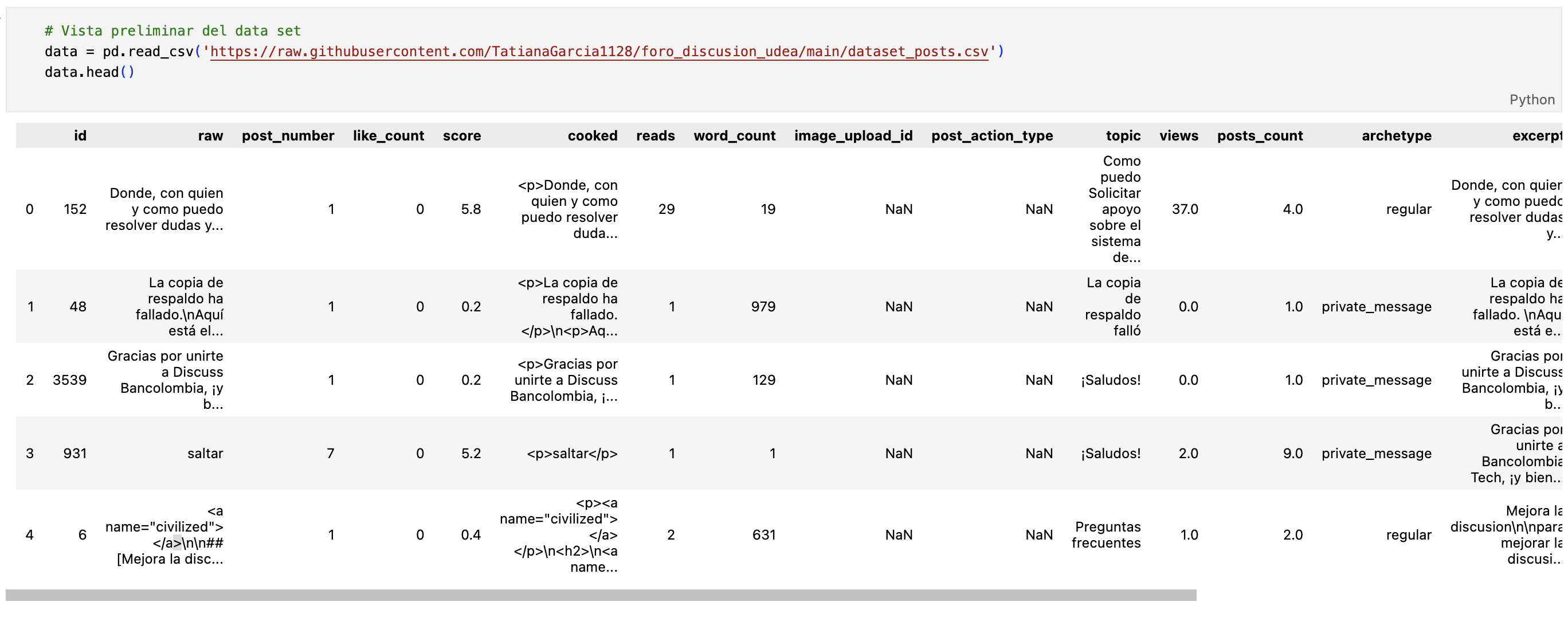


Imagen 6: Muestra dataset post

A continuación, se describen cada una de las columnas que hacen parte del dataset:

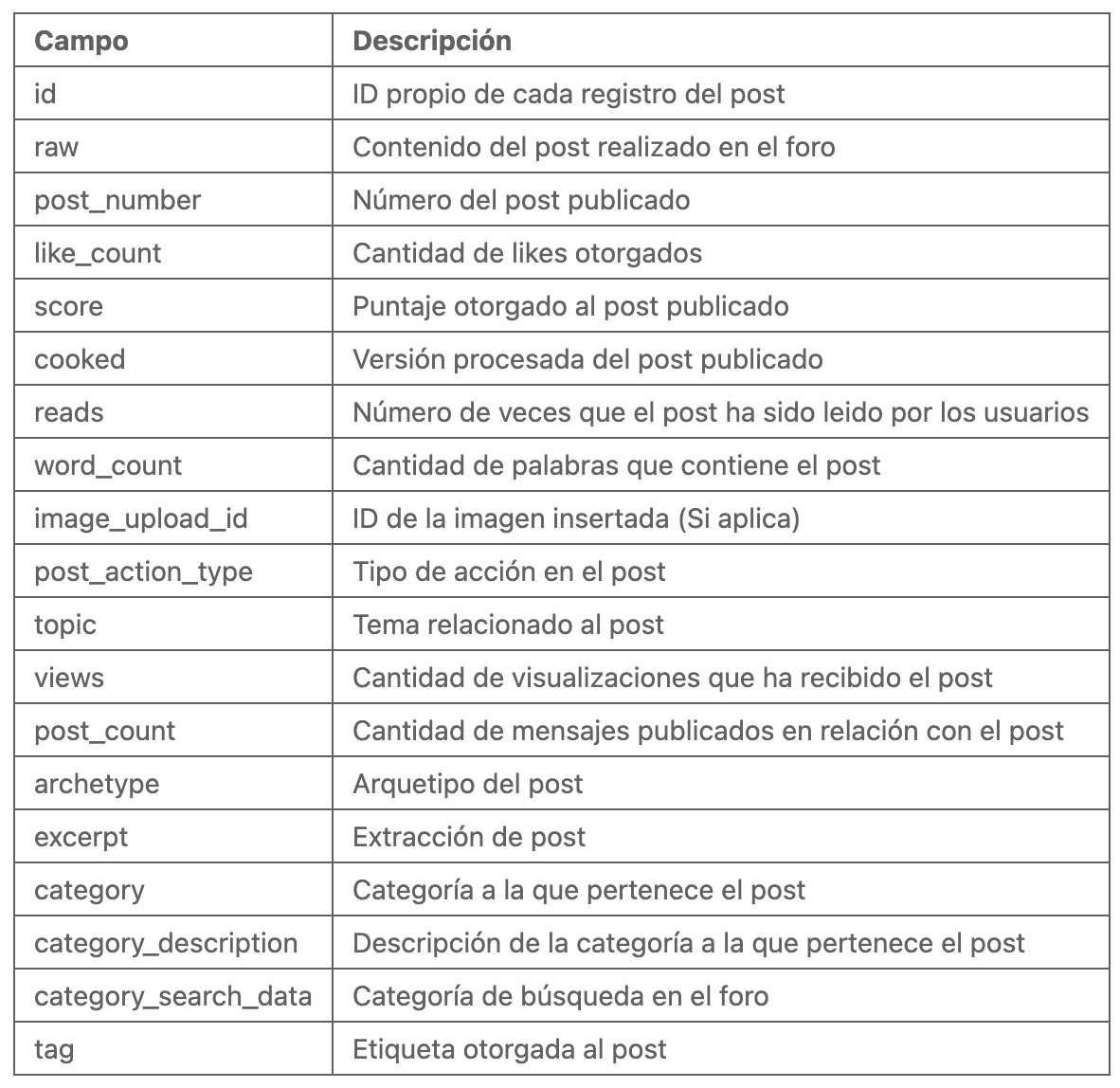


Imagen 7: Descripción de datos dataset posts

El dataset contiene 10547 instancias y 19 columnas, de las cuales 5 tienen valores tipo entero (int64), 4 de tipo decimal (float64) y las 10 restantes son tipo object. Presenta las siguientes columnas con valores nulos 'image\_upload\_id', 'post\_action\_type', 'raw', 'topic', 'views', 'posts\_count', 'archetype', 'excerpt', 'category', 'category\_description', 'category\_search\_data' y 'tag' y en total el peso del dataset es de 1.5+ MB; es decir, que es fácil de procesar.

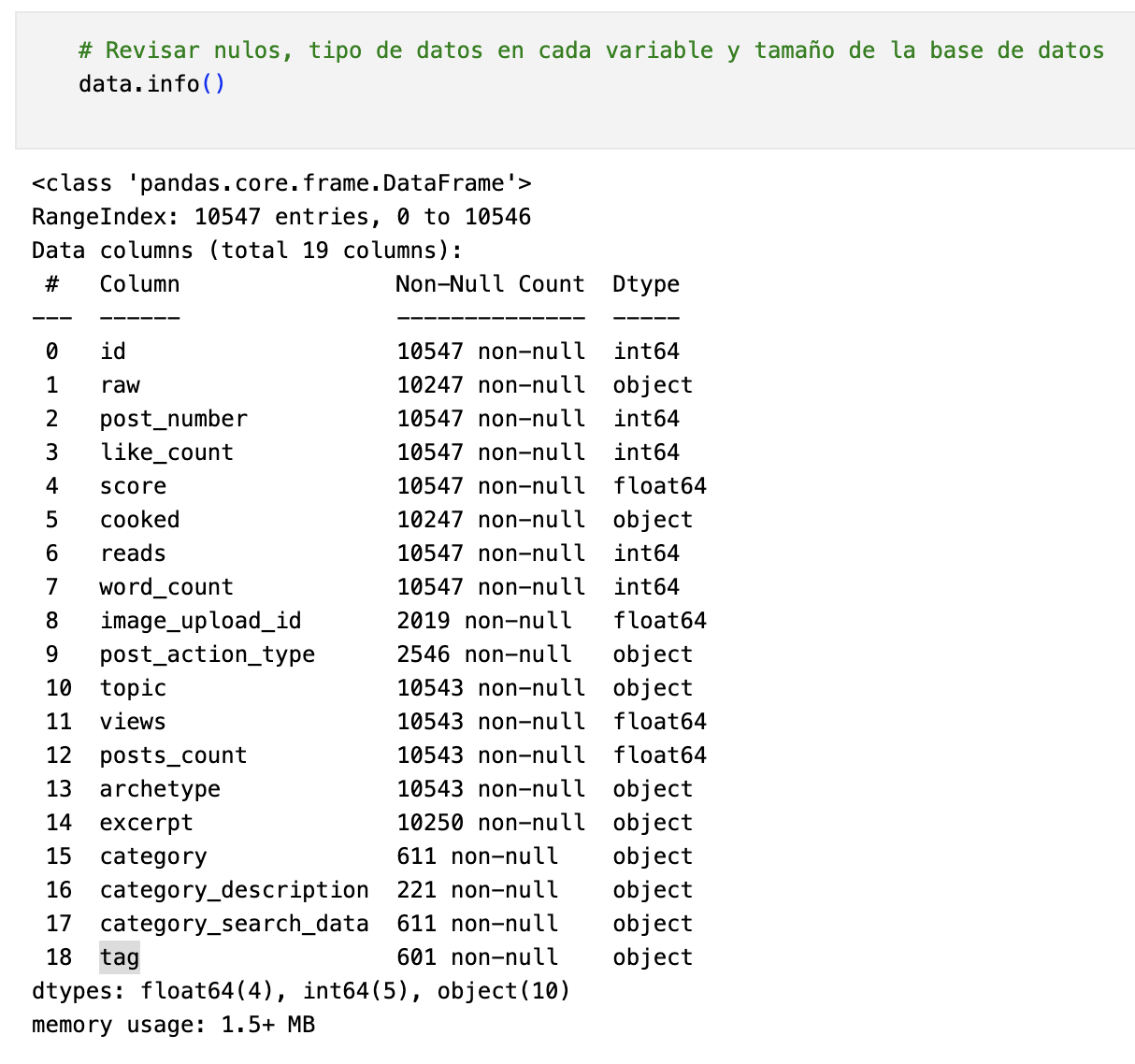


Imagen 8: Descripción detallada del dataset

El dataset presenta escalas diferentes por lo que posiblemente se requiera una estandarización de la información en pasos posteriores.

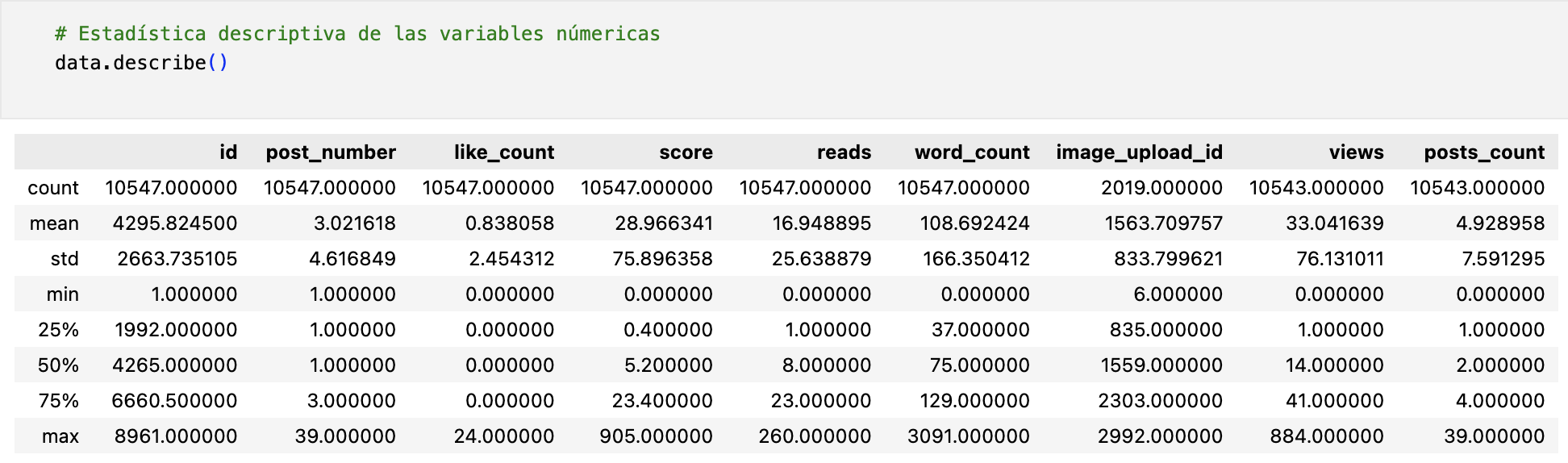


Imagen 9: Información descriptiva del dataset

## Datsets

Para la segmentación de usuarios se realiza una limpieza de los datos evaluando los siguientes puntos:

* Eliminación de valores nulos
* Eliminación de valores duplicados
* Eliminación de columnas que no aportarían al modelo
* Reemplazar los valores nulos de la variable ‘category\_description’

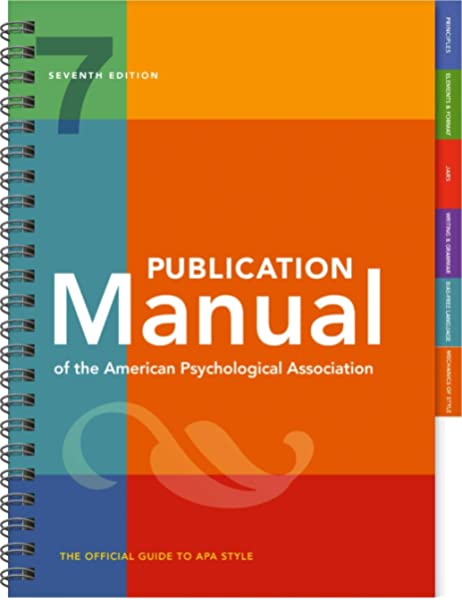
Finalmente, de 24 columnas se reduce a 21.

Para la evaluación de calidad de las respuestas proporcionadas por los usuarios en el foro de discusión se realiza la limpieza de datos mediante las siguientes tareas:

* Eliminación de valores nulos
* Eliminación de valores duplicados
* Eliminación de columnas que no aportarían al modelo
* Reemplazar los valores nulos en las columnas correspondientes.
* Limpieza de etiquetas html.
* Creación de bolsa de palabras y tokenización

Finalmente, de 10.547 instancias y 19 columnas quedan 4.472 y 16 respectivamente.

# Figura 1 Portada Normas APA séptima edición 2020 en inglés



*Nota.* Fuente https://bit.ly/2IyrZao (American Psychological Association, 2020).

**Figura 2**  
Logo Universidad de Antioquia

****

*Nota.* Fuente http:/www.udea.edu.co

# Referencias

[1] Gil-Vera, V. D., Seguro-Gallego, C. (2022). Machine Learning Aplicado al Análisis del Rendimiento de Desarrollos de Software. Fuente. Recuperado de <https://www.redalyc.org/journal/6078/607870799010/607870799010.pdf>.

[2] León Guzmán, E. (s.f.). Métricas para la validación de Clustering. Recuperado de

<https://disi.unal.edu.co/~eleonguz/cursos/mda/presentaciones/validacion_Clustering.pdf>.