María Eugenia Curi
Lucía Osorio

# Tarea 1 - Introducción a la Ciencia de Datos

## Introducción

A continuación, se ofrece una descripción de los datos explorados, organizados en cuatro tablas principales: works, characters, chapters y paragraphs.

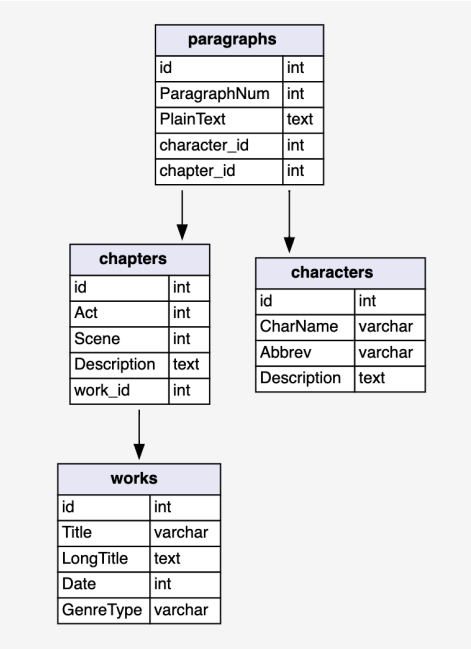


Fig.1 tablas de la base de datos Shakespeare

La tabla **works** almacena información sobre las obras de Shakespeare, incluyendo el título breve y completo, la fecha de creación y el género. Contiene un total de 43 obras. Un ejemplo es la comedia "Twelfth Night", cuyo título completo es "Twelfth Night, Or What You Will", y data del año 1599.

En la tabla **chapters**, que contiene 945 registros, se detallan los capítulos de cada obra, vinculados a través de la columna *work_id*. Esta tabla desglosa la estructura de cada obra en actos y escenas, además de incluir una descripción del lugar de los eventos. Por ejemplo, "Twelfth Night" está dividida en 5 actos, siendo el primer acto compuesto por 5 escenas.

La tabla **paragraphs** registra el texto de cada capítulo, vinculado a los personajes y a los capítulos específicos. Siguiendo con el ejemplo anterior, el primer acto y la primera escena de "Twelfth Night" contienen 8 párrafos.

Por otro lado, la tabla **characters** recopila información sobre los personajes de las obras, incluyendo su nombre abreviado, completo y una descripción. Hay un total de 1266 personajes, uno de ellos es Orsino, también conocido como Duke Orsino, el Duque de Illyria, quien aparece en el primer capítulo de "Twelfth Night".

En cuanto a la calidad de los datos, en la tabla **characters** aparecen como nulos cinco registros de la columna **Abrev** (Players, Earl of Kent, John of Lancaster, Senator y Earl of Surrey), y 646 registros en la columna **Description**.

Durante la exploración de datos, en búsqueda del personaje con más párrafos, se identificaron acotaciones y voces poéticas atribuidas a Shakespeare, que si bien aparecen en esta tabla se excluyen de esta categoría por no cumplir con el criterio de personaje. Al finalizar este proceso, se encontró que el personaje con más párrafos es Falstaff, con 471 de éstos últimos.

## Exploración de obras

Al explorar la fecha de publicaciones de Shakespeare se obtiene la siguiente gráfica (fig. 2), que presenta la cantidad de obras por año.

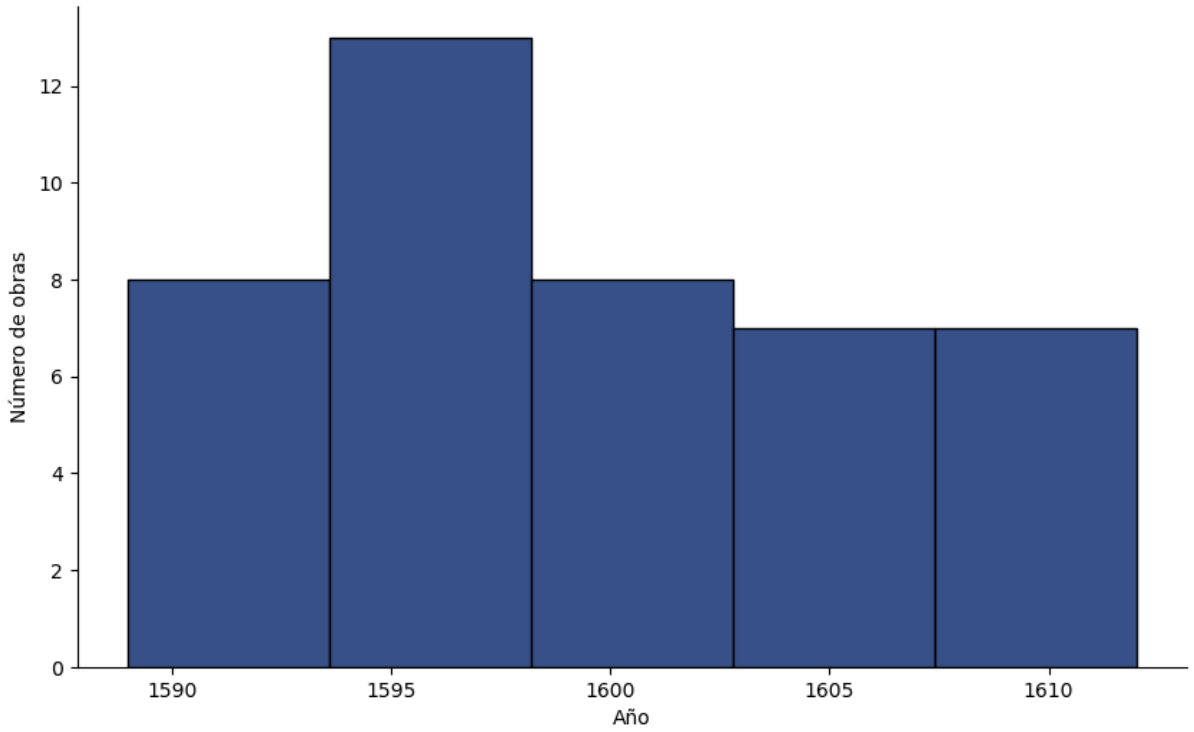


Fig. 2.a Histograma de producción de obras a lo largo de los años

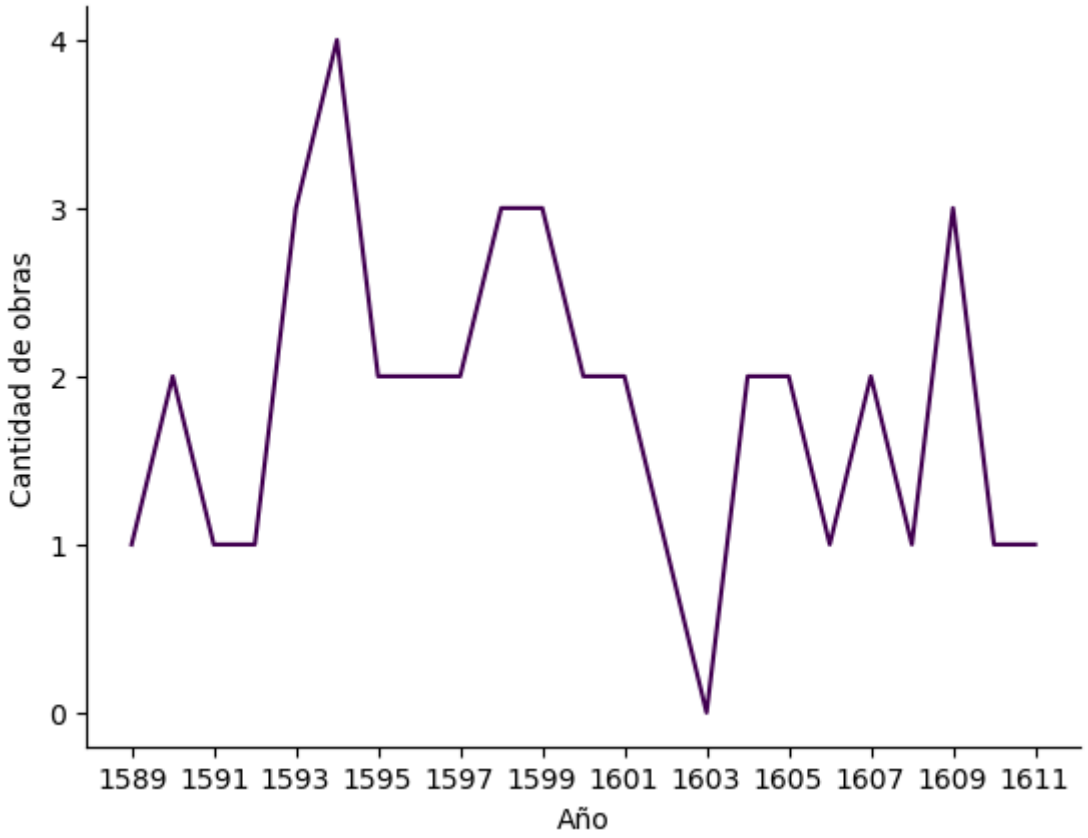


Fig. 2.b Producción de obras a lo largo de los años

Se ha observado un incremento notable en la publicación de obras de Shakespeare, que abarcan desde el año 1589 hasta 1611. El año con mayor obras fue 1594, durante el cual se publicaron cuatro de éstas, culminando un período de crecimiento iniciado en 1592. Sin embargo, entre los años 1601 y 1603 se registró un descenso en la publicación, siendo 1603 un año en el que no se registraron obras.

En la figura 3 se presenta un análisis exploratorio sobre la cantidad de obras por género .

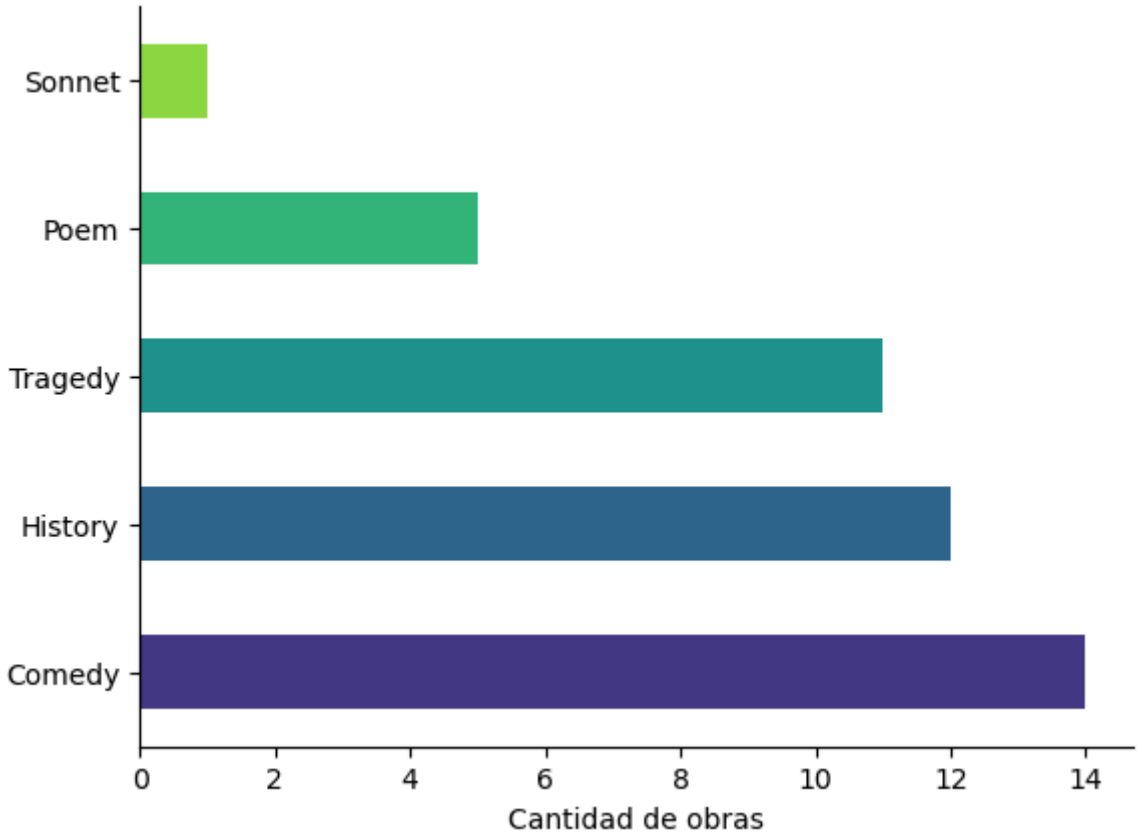


Fig. 3 Cantidad de obras por género

Se observa que la colección incluye un soneto y cinco poemas. Sin embargo, los géneros más representados son la tragedia, la historia y la comedia, con este último género constituyendo el **33%** del total de las obras.

Se visualiza el género más publicado por año, junto con la cantidad de publicaciones del mismo (fig. 4). Se puede observar que en doce años el género comedia obtiene la mayor frecuencia, siendo el género más publicado a lo largo de los años. Salvo en seis ocasiones que lo hace la historia, en tres que se publican tragedias mayormente y en un año que fue el género poema el más frecuente. Resulta interesante visualizar que el soneto no es publicado con mayor frecuencia en ninguno de los años.

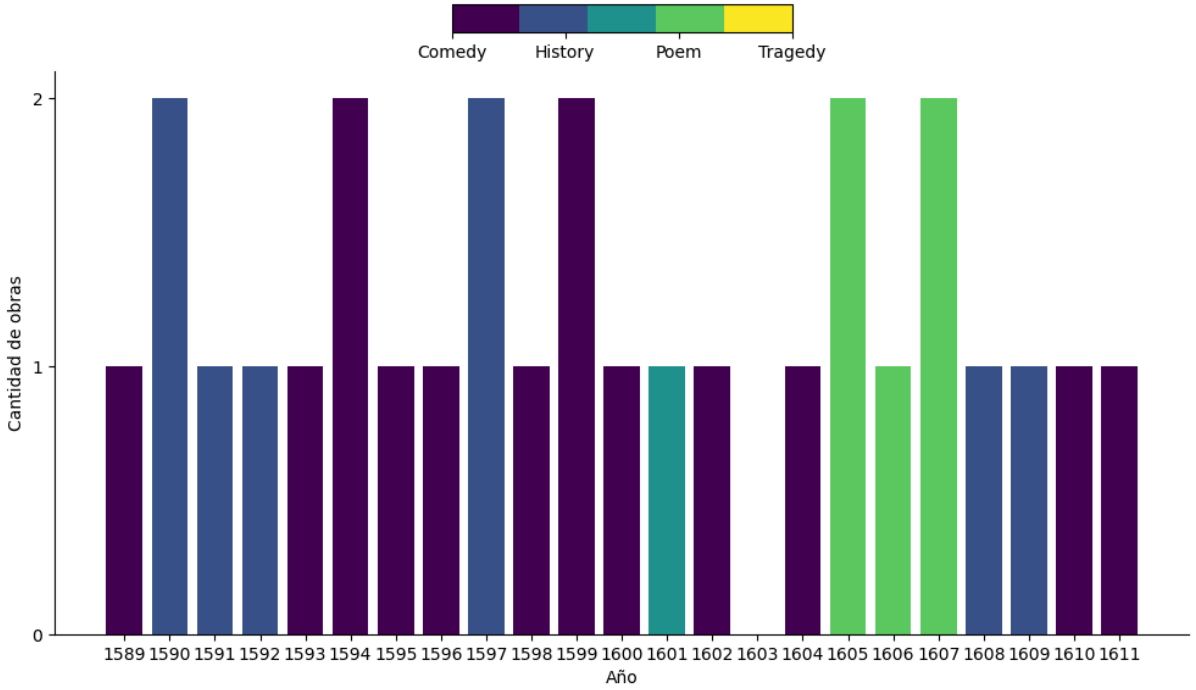


Fig. 4 Géneros más producidos por año

Para generar visualizaciones útiles, nos centraremos en estos tres géneros predominantes.


Fig 5.a. Obras del género comedia a lo largo del tiempo

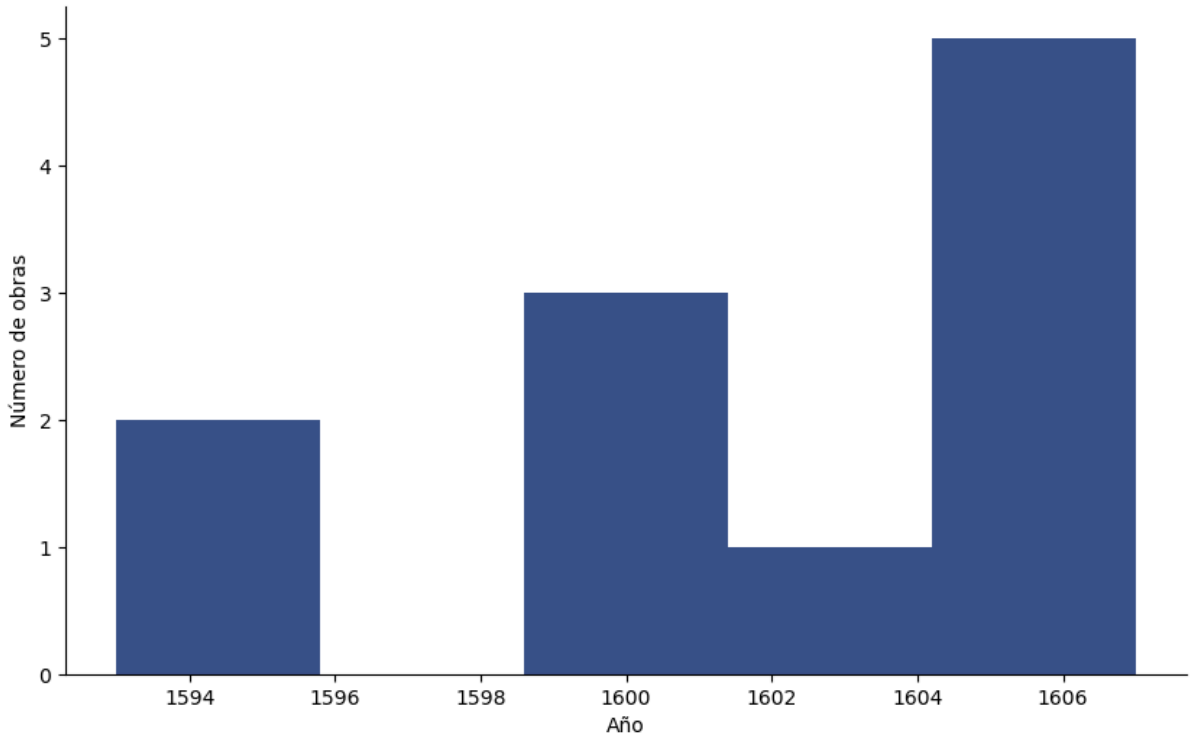


Fig 5.b. Obras del género tragedia a lo largo del tiempo

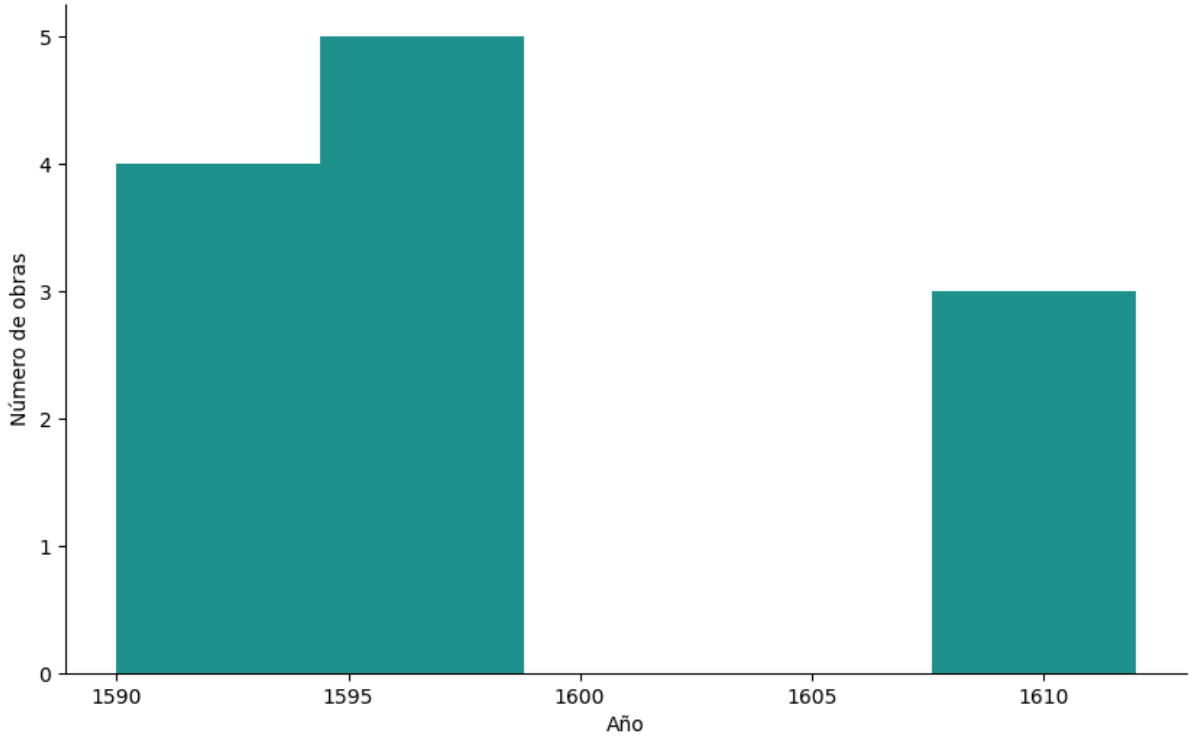


Fig. 5.c Obras del género historia a lo largo del tiempo

Como se muestra en la figura 5, el género de la tragedia experimentó un aumento significativo después de 1603, año hasta el cual la máxima publicación era de dos tragedias, alcanzando su máximo de producción en el período alrededor del 1600. En

contraste, las obras históricas dominaron la producción de Shakespeare hasta el año 1600, momento a partir del cual su enfoque en este género disminuyó notablemente. En cuanto a las comedias, Shakespeare mantuvo una producción constante de este género a lo largo de casi todos los períodos, con la excepción de dos momentos específicos, tal como se indica en la gráfica.

## Exploración de palabras

Para poder identificar correctamente las palabras utilizadas en las obras fue necesario realizar una limpieza del texto. En primer lugar se exploraron los primeros 10 párrafos del df_paragraphs donde identificamos los signos de puntuación que aparecían y los agregamos a la lista de en la función clean_text. Los signos de puntuación que debieron eliminarse fueron: ”;","]",".","?",":","!","\"","\'","(",")","--"

Uno de los signos más utilizado era la comilla simple (’) debido al alto uso del apóstrofe en el idioma inglés, en este caso se filtró en la tabla df_words aquellas palabras que contuvieran este símbolo, con algunas salvedades. Por ejemplo, de las palabras más usadas se decidió separar: I’ll por I will - there’s por there is - that's por that is - who's por who is. Al comienzo de la exploración una posibilidad era modificar ‘s por is, pero eso no era posible ya que en algunos casos se hacía referencia a la pertenencia por ejemplo: “father's court”. Por eso se decidió hacer los cambios específicos descritos anteriormente. Otro ejemplo que no pudo cambiarse de forma masiva fue la contracción “ ‘d” debido a que en algunas ocasiones era la forma abreviada de “had”  y en otras de “would”.

### Palabras más utilizadas

Una vez realizada esta limpieza, se visualizan las 10 palabras más utilizadas  en el texto (fig. 6).

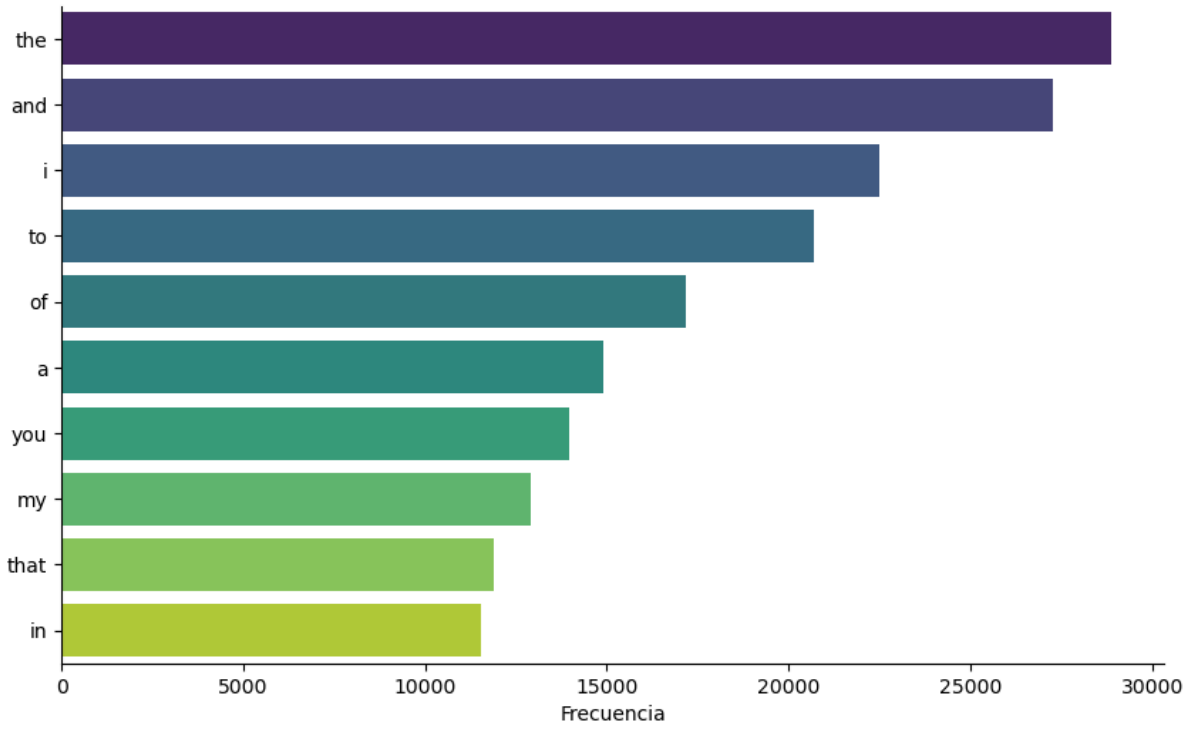


Fig. 6 10 Palabras más utilizadas en las obras

Las palabras más utilizadas son palabras comunes del idioma inglés lo cual no nos proporciona ninguna información adicional sobre la obra de Shakespeare y su contenido. Si deseamos tener algún tipo de percepción más profunda sobre la naturaleza de la misma podríamos realizar una lista de las palabras más comunes del idioma inglés (preposiciones, pronombres, etc), filtrar nuestro dataframe de palabras usando esta lista y volver a graficar. Adicionalmente podríamos realizar gráficos comparando la frecuencia de ciertos grupos de palabras, un ejemplo sería evaluar la frecuencia con que se utilizan palabras relacionadas a diversas emociones y compararlo con el género de la obra.

Si se quisiera presentar la cantidad de usos de una palabra por personaje se podría generar una tabla con la cantidad de veces que repite esa palabra cada personaje.

## Personajes con mayor cantidad de palabras

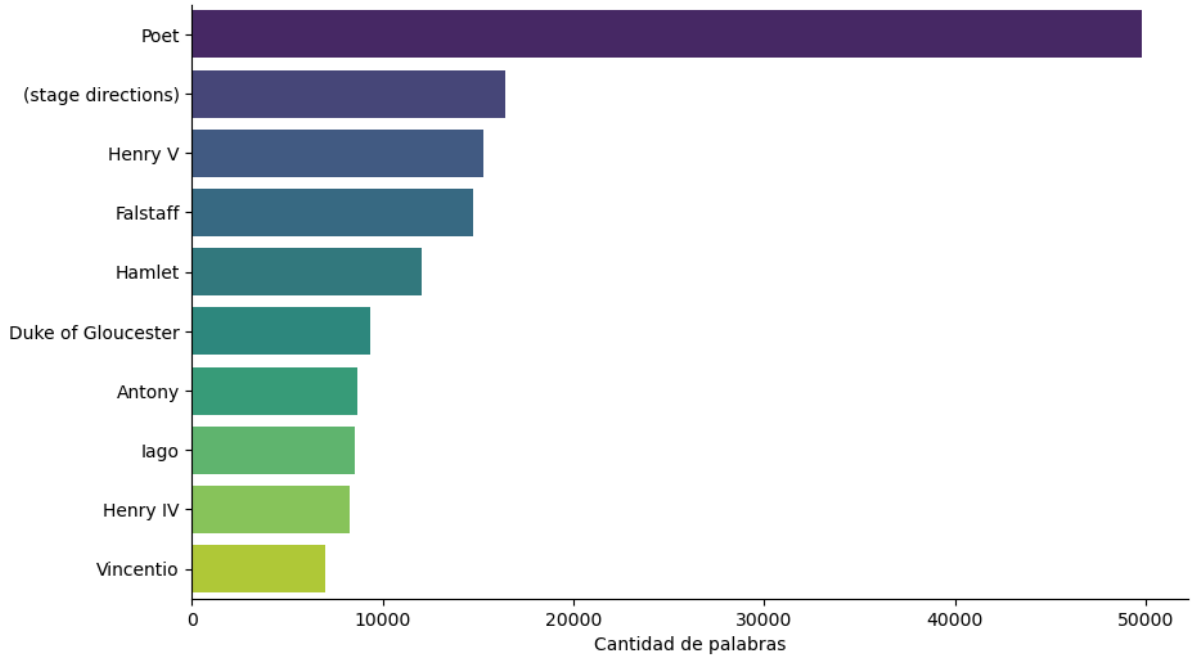


Fig. 7 Cantidad de palabras por personaje

Para identificar a los personajes que tengan mayor cantidad de palabras fue necesario filtrar “Poet” y “(stage directions)” como nombres de personajes ya que no se los considera dentro de esta categoría (fig. 7). En la figura 8 se presentan los 10 personajes con mayor cantidad de palabras, siendo Henry V el personaje con mayor cantidad de palabras.
Se escogió utilizar los personajes en el eje de las Y para una mejor legibilidad de los datos.

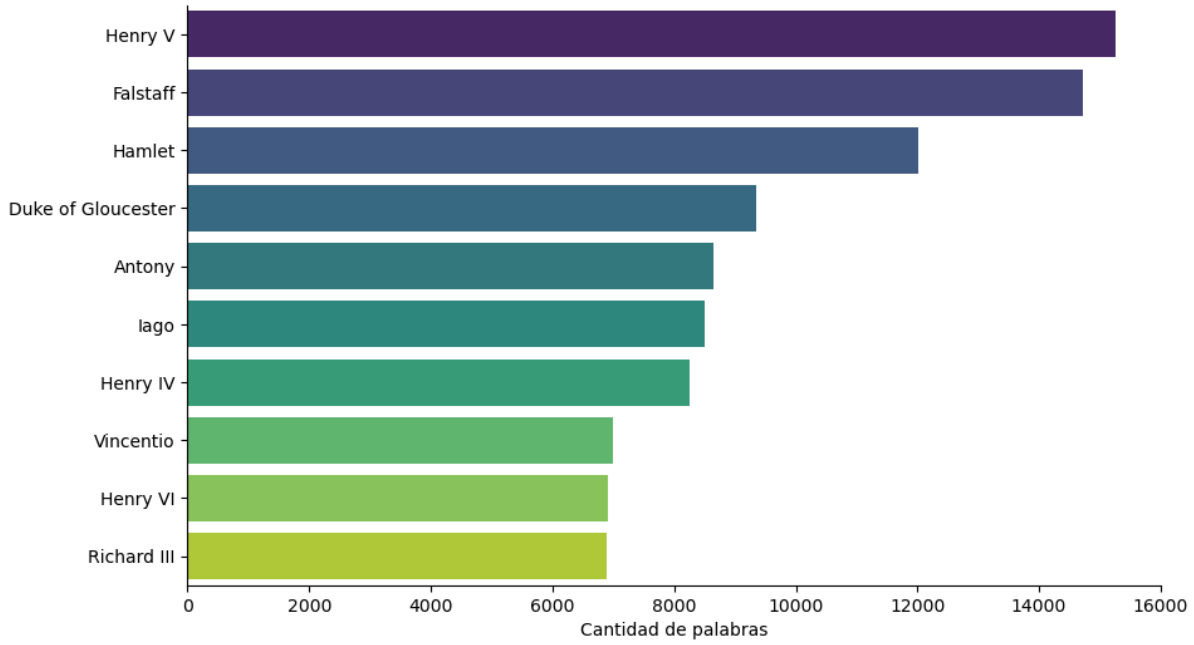


Fig. 8 Cantidad de palabras por personaje corregido por personajes reales

## Trabajo a futuro

Las tablas permiten responder aún mayor cantidad de preguntas. Respecto a los géneros, sería interesante explorar aquellos con mayor cantidad de palabras o capítulos, ¿existen patrones? Lo mismo podría realizarse con las obras y los personajes, ¿qué obra tiene mayor cantidad de personajes?

Respecto a la visualización respecto al año, se podría presentar la palabra más frecuente por año.