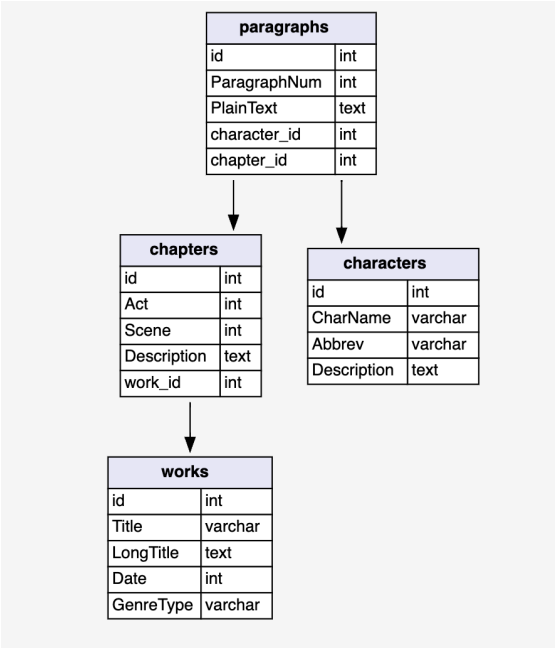
Introducción a la ciencia de datos

Ejercicio 1

A. *Cargar todas las tablas disponibles*
   a. *Comente la función de cada tabla y la relación entre ellas.*
   b. *Reporte si existen datos faltantes en algún campo, o cualquier otro problema de calidad de datos que encuentre. En particular, analice la cantidad de párrafos por personaje. ¿Cuál es el personaje con más párrafos?*

Modelo de Datos



El modelo de datos está compuesto por cuatro tablas:

1. **Paragraphs**: en esta tabla vamos a tener cada línea escrita en los trabajos de Shakespeare con el ID del personaje (character_id) que la dice y el ID del capítulo (chapter_id) en el que se encuentra dicha línea. La tabla tiene un id único por párrafo más un identificador de párrafo interno por capítulo (paragraphNum).
2. **Chapters**: en esta tabla se encuentra una descripción de los capítulos de cada obra de Shakespeare. La tabla tiene un ID único por chapter, además de indicar el número de acto (Act) dentro de la obra sumado al número de escena (Scene) dentro de ese acto. Por último agrega una descripción(Description) sobre donde se desarrolla dicho capítulo y el id de la obra a la cual pertenece el capítulo en cuestión (work_id).

3. **Works**: en esta tabla se encuentran todas las obras de Shakespeare, con su título corto (Title), titulo largo (LongTitle), la fecha de publicación (Date) y el género (GenreType). El campo ID representa un identificador único de cada obra.
4. **Characters**: en esta tabla se encuentran todos los personajes que aparecen en las obras de Shakespeare, con su nombre (CharName), su abreviación (Abbrev) y una descripción del mismo (Description). Cada personaje tiene su ID único en la tabla.

Con respecto a las relaciones entre las tablas podemos indicar lo siguiente:

1. Cada obra (work - id) tiene 1 o N capítulos (chapter - id).
2. Cada capitulo (chapter - id) corresponde a 1 obra (work_id).
3. Cada capitulo (chapter - id) tiene 1 o N párrafos (paragraph - id).
4. Cada parrafo (paragraph - id) corresponde a 1 capítulo (chapter - id).
5. Cada parrafo (paragraph - id) corresponde a 1 personaje (character - id).
6. Cada personaje (character - id) tiene 0 o N párrafos (paragraph - id).

Calidad de datos

Characters:

Filas Totales: 1266

Datos Faltantes:
id : 0 (0.00%)
CharName : 0 (0.00%)
Abbrev : 5 (0.39%)
Description : 646 (51.03%)

Tipos de datos:

id	int64
CharName	object
Abbrev	object
Description	object

Filas Duplicadas: 0

Más allá de los datos faltantes en la tabla de personajes, encontré otro problema que es que parecen estar duplicados dichos personajes, es decir un mismo personaje tiene más de un ID único que lo representa, un ejemplo que me llamó la atención fue el “príncipe de Francia”:

635	King of France	KING	NaN
636	King of France	France	NaN

Donde podemos notar que es muy probable que sea el mismo personaje. Otros ejemplos son los personajes denominados como Gentleman, Lord, Messenger, Officer, etc que también aparecen con diferentes ID sin poder tener la precisión si son el mismo o no.

Works:

Filas Totals: 43

Datos Faltantes:
id : 0 (0.00%)
Title : 0 (0.00%)
LongTitle : 0 (0.00%)
Date : 0 (0.00%)
GenreType : 0 (0.00%)

Tipos de datos:

id	int64
Title	object
LongTitle	object
Date	int64
GenreType	object

Filas Duplicadas: 0

Chapters:

Filas Totals: 945

Datos Faltantes:
id : 0 (0.00%)
Act : 0 (0.00%)
Scene : 0 (0.00%)
Description : 0 (0.00%)
work_id : 0 (0.00%)

Tipos de datos:

id	int64
Act	int64
Scene	int64
Description	object
work_id	int64

Filas Duplicadas: 0

Paragraphs:

Filas Totals: 35465

Datos Faltantes:
id : 0 (0.00%)
ParagraphNum : 0 (0.00%)
PlainText : 0 (0.00%)

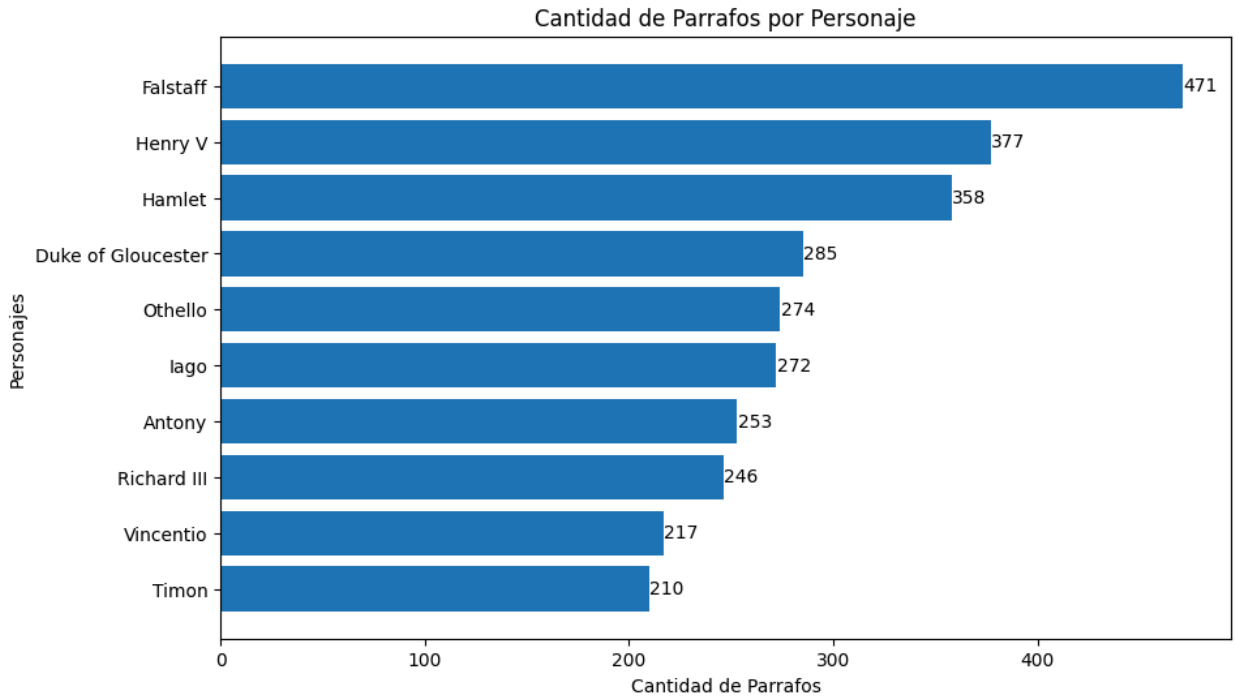
character_id : 0 (0.00%)
chapter_id : 0 (0.00%)

Tipos de datos:

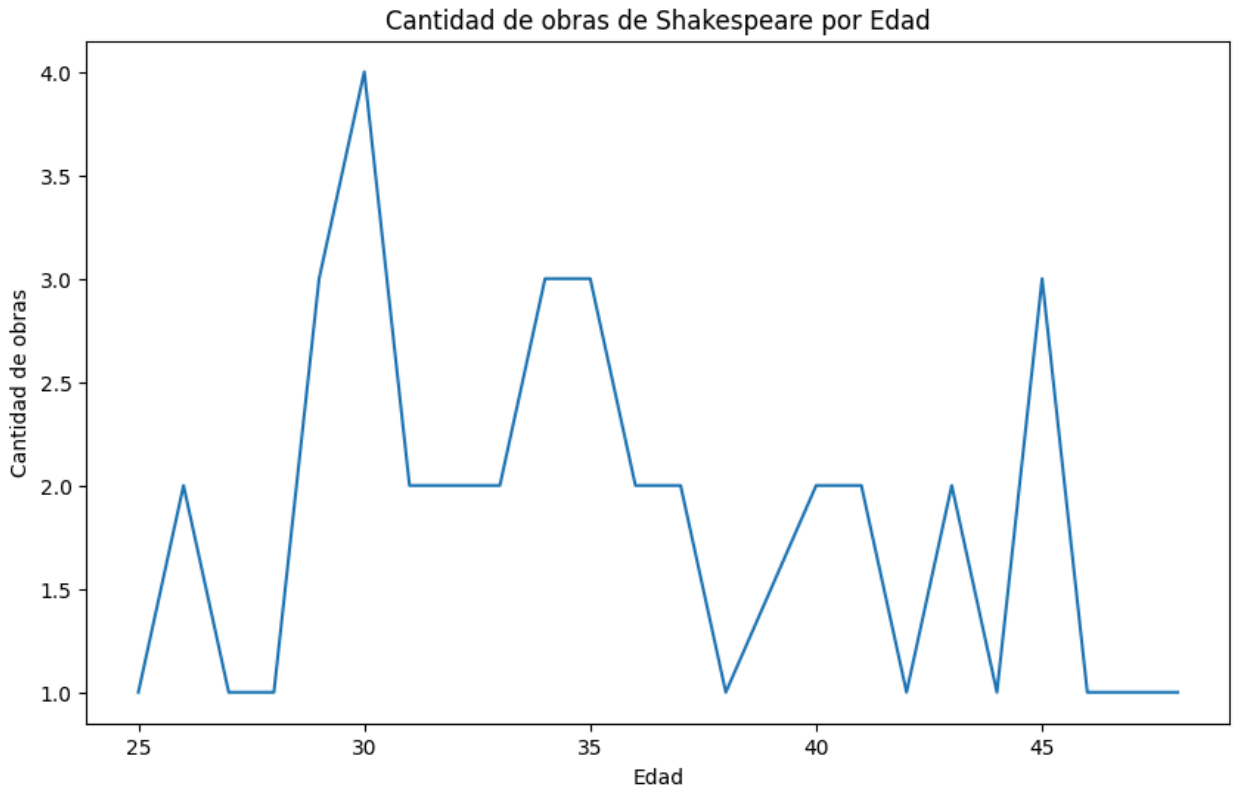
id	int64
ParagraphNum	int64
PlainText	object
character_id	int64
chapter_id	int64

Filas Duplicadas: 0

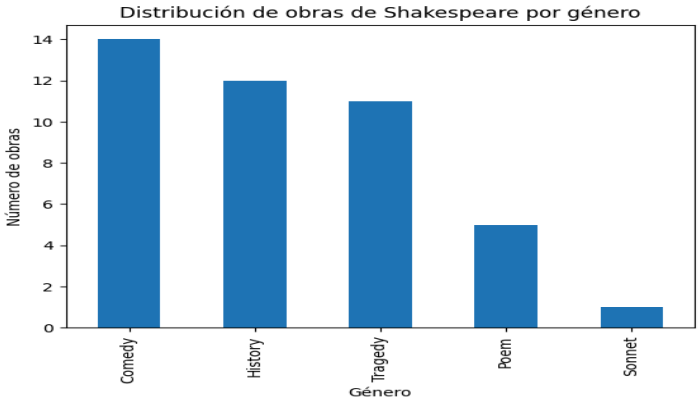
Con respecto al personaje con más párrafos, quitando los valores “(stage directions)” y “Poet” (*the voice of Shakespeare’s poetry)* que entendemos que no son personajes per se en sus obras, podemos decir que el Falstaff es el personaje que más párrafos tiene en las obras de Shakespeare y luego lo siguen Henry V y Hamlet.



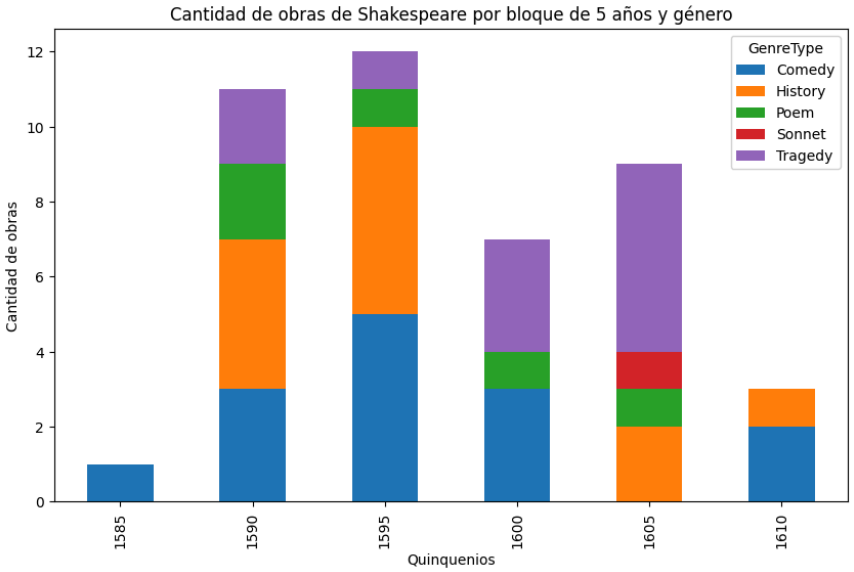
B. *Genere una gráfica que permita visualizar la obra de Shakespeare a lo largo de los años. Por ejemplo, tomando períodos de algunos años y mostrando la cantidad de obras escritas para esos períodos. Comente si se observan tendencias (o no) a lo largo del tiempo, por ejemplo respecto a su producción, o los géneros sobre los que escribió. No realizar análisis estadísticos, solamente generar visualizaciones exploratorias.*



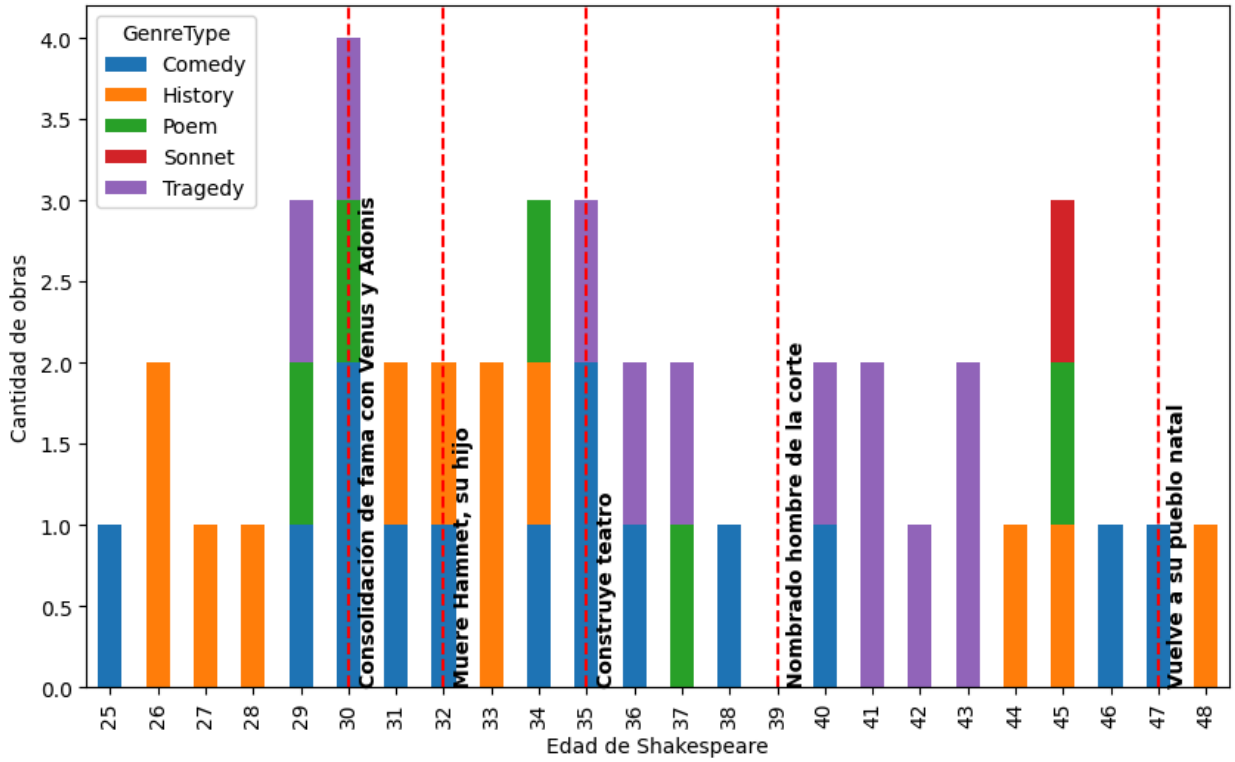
Tomando en cuenta que William nació en 1564, podemos apreciar que publicó su primera obra a los 25 años y luego lo hizo de forma consistente todos los años siguientes, publicando al menos 1 obra al año (menos en 1603). Su pico de productividad lo tuvo a sus 30 donde publicó 4 obras en un mismo año calendario. Con respecto a sus géneros más populares, podemos ver que durante su vida lo que más escribió fue Comedia, seguido de Historia y Tragedia.



Pero si abrimos sus obras por género y por quinquenio de publicación podemos notar que la comedia estuvo tanto al inicio de su vida como en su fin, como así la historia pero parece que la tragedia fue algo que desarrolló en su momento de mayor producción pero que abandonó al final.



Si abrimos el gráfico por edad podemos ver la importancia que tuvo la tragedia en la mitad de su vida llegando a publicar 6 obras en 4 años ininterrumpidos (entre los 40 y los 43), lo cual no tiene precedente con otros de sus géneros. Por último cabe destacar que la publicaciones de sonetos parece haber sido algo de una sola vez a sus 45 años.



C) Conteo de palabras. Comente todas las transformaciones de texto que haya agregado y justifique.

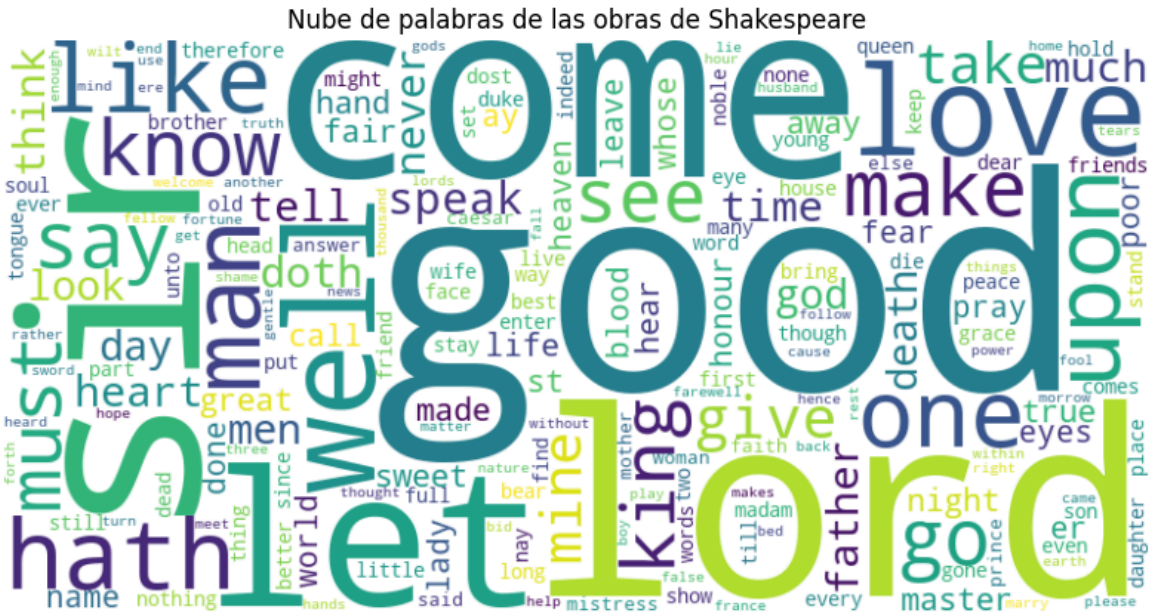
Antes de realizar el conteo de las palabras de las obras de shakespeare se realizaron las siguientes reglas de limpieza:

1. Eliminación de expresiones teatrales que se encontraban en el texto, como por ejemplo [Exeunt], buscando todo texto encerrado entre corchetes y eliminandolo.
2. Eliminación de los caracteres especiales de saltos de línea, tabulaciones, etc que puede contener el texto en la base de datos.
3. Eliminación de los signos de puntuación usando la constante de python *string.punctuation*.
4. Eliminación de los dígitos que se encuentran en el texto.
5. Eliminación de las stopwords o palabras vacías, que son palabras que se consideran de bajo valor semántico en un texto. Tuvimos que agregar stopwords tanto en inglés moderno como en Middle English ya que los textos tienen palabras que hoy no son consideradas vacías ya que se dejaron de usar. Por ejemplo, thou.

Con las transformaciones anteriores, lo que buscamos es dejar el texto lo más limpio posible de palabras con valor semántico y significado en las obras de Shakespeare, y así lograr un análisis que se centre principalmente en las palabras más relevantes de su obra.

Ejercicio 2

A) Realice una visualización que permita comparar las palabras mas frecuentes, considerando toda su obra.



Sin necesidad de implementarlo, proponga ideas para modificar esta visualización con el fin de encontrar diferencias entre géneros y personajes.

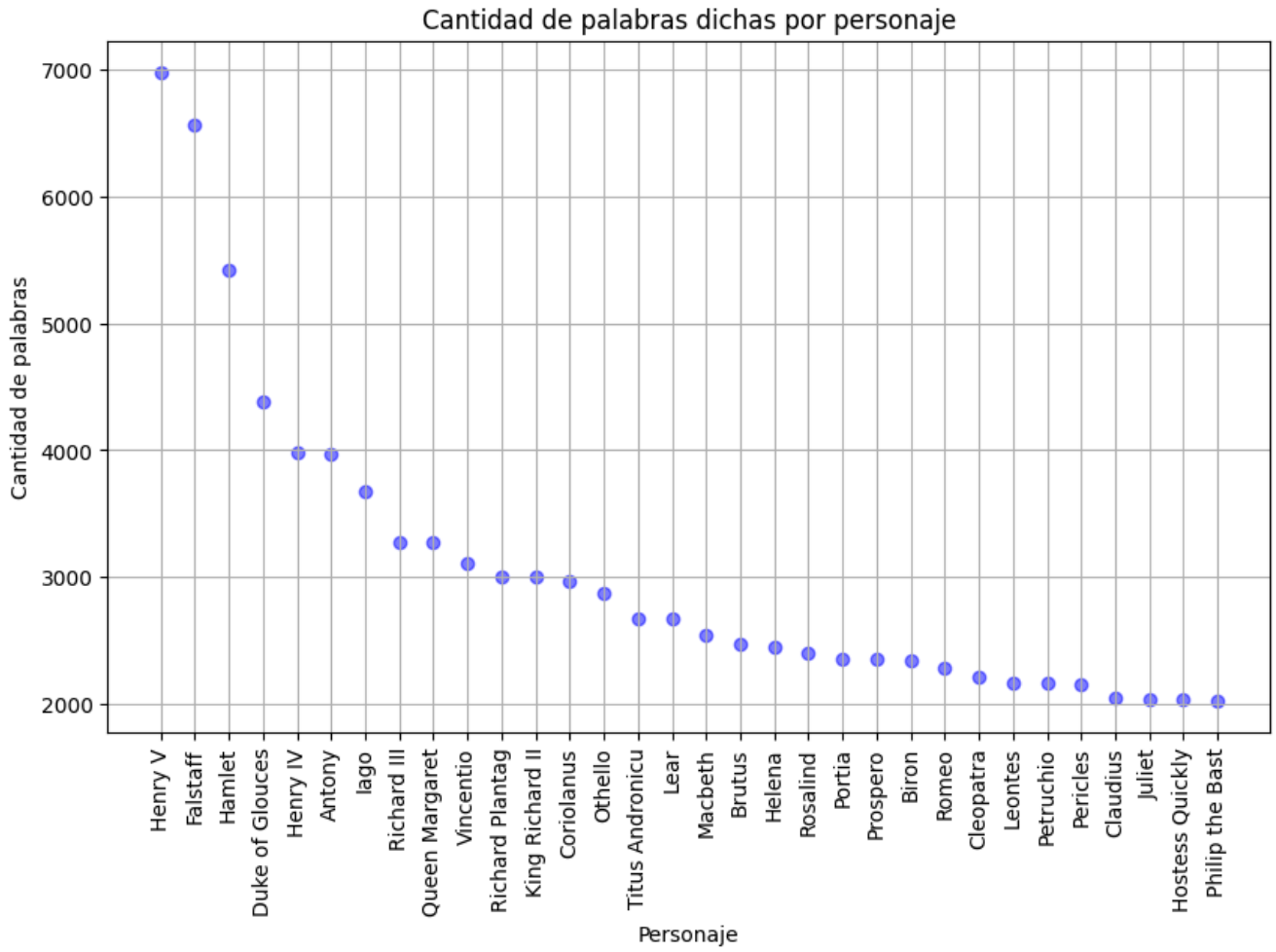
Lo que me parecería interesante para destacar géneros o personajes en la visualización anterior sería:

- Para diferenciar géneros, usar diferentes colores por género, entonces si bien la estructura de la gráfica se mantendrá, a partir de los colores fácilmente podría identificar si las palabras más repetidas son dichas por hombres, mujeres o cualquier otro género.
- La misma técnica de usar colores o tipografías , podría aplicarlo a personajes aunque en este caso la cardinalidad de las variantes aumentan y no sería tan claro de analizar a simple vista.

B) Corra el código que permite encontrar los personajes con mayor cantidad de palabras. En caso de encontrar algún problema luego de realizar la visualización, comente a que se debe y proponga formas de resolverlo.

El primer problema que uno se encuentra cuando corre el codigo para ver los personajes con más palabras dichas es que aparece en los primeros puestos personajes como Poet o Stage Directions que si bien están en la tabla de personajes , no son personalidades representativas por si solas en las obras si no que son conceptos más genéricos. Mi manera de solucionar dicho tema fue el siguiente:

1. Eliminando las expresiones teatrales, es decir los párrafos que estaban entre corchetes, se eliminó solo el personaje “Stage Directions”.
2. Por otro lado, mirando la tabla de personajes puedo notar que muchísimos personajes genéricos de dicha tabla tienen description NaN, entonces tomé la decisión de filtrar todos esos personajes del análisis.
3. Por ultimo, elimine Poet “the voice of Shakespeare’s poetry” manualmente ya que no me era relevante.



C) Proponga preguntas que se podrían intentar responder a partir de estos datos, y mencione posibles caminos para responderlas.

Las primeras dudas o preguntas que me surgen de lo visto anteriormente es saber si los personajes a los que se le adjudican más palabras es porque realmente tuvieron una presencia preponderante en las obras de shakespeare o hay excepciones de personajes que son protagonista solamente en una de las obras más extensas de William y por eso lidera el ranking.
Otra de las dudas es si hay una tendencia a textos más largos o más cortos según el género del libro, como también saber cuan preponderante es la presencia de personajes con títulos

nobiliarios en sus obras, ya que en los primeros puestos ya se pueden ver nombres como Queen, King, Duke, etc.

Los caminos para responder dichas preguntas salen de agregar nuevas dimensiones a los análisis anteriores. Dimensiones como género del libro, clustering de personajes según si en su nombre tienen títulos nobiliarios o no, largo total del texto con relación a la cantidad de palabras dichas por cada personaje. Creo que se puede plantear diferentes matices a análisis previos sin agregar nuevos dataset a los existentes.