

# Análisis exploratorio de datos

F.R.I.E.N.D.S

Aplicando la ciencia de datos para resolver grandes cuestiones de la humanidad, en este capítulo...

¿Quién es el auténtico protagonista de 'Friends'?



# Aplicando la ciencia de datos para resolver grandes cuestiones de la humanidad, en este capítulo...

Razones para haber elegido este tema:

- ◆ Da igual los años que pasen que no envejece
- ◆ Me rio hasta llorar aunque haya visto la misma escena 100 veces
- ◆ Sabes que siempre estará ahí para acompañarte en un día chungo
- ◆ Es la mejor serie del universo y siempre lo será
- ◆ Y otras tantas razones que cada uno tendrá en su ❤️



## Origen de los datos

En este estudio, hemos utilizado tres datasets provenientes de la web [kaggle](#). De ellos, destacamos la información más relevante que nos aportan:

- ◆ Dataset completo con la transcripción en texto de todos los diálogos de la serie (si, es real 🤖).
- ◆ Dataset completo con todas las temporadas y capítulos y su valoración en IMDB.
- ◆ Dataset completo con todas las temporadas y capítulos y los datos de visionado en EEUU.



## ¿De qué datos hemos partido?

Para que nos hagamos una idea de la magnitud de los datos:

- ◆ La serie tiene 10 temporadas y 228 capítulos
- ◆ Hay un total de 459 personajes que aparecen a lo largo de toda la serie
- ◆ Se dicen en total, 61.256 frases
- ◆ El estudio gira alrededor de estos 6 personajes principales ↘

CHANDLER



JOEY



PHOEBE



RACHEL



ROSS



MONICA



## Limpieza de los datos

Debido a que principalmente hemos trabajado con datos en formato texto, la parte dedicada a su limpieza ha tenido bastante peso en relación al tiempo empleado en el estudio completo.

La columna de personajes se ha limpiado de 940 valores que parecían “únicos” a 459 valores REALMENTE únicos.

Hemos salvado todos los ‘\xa0joey’, todos los ‘rachel(crying)’, muchos ‘rach’, ‘mnca’, ‘chan’, ‘phoe’, montones de ‘ross to monica’ y otros tantos errores tipográficos muy confusos.

Las 60k líneas correspondientes a los guiones de cada capítulo, se han limpiado igualmente, eliminando caracteres raros y más importante, eliminando todas las acotaciones (entre paréntesis) que invalidaban nuestros resultados.



# Hipótesis

¿Quién es el protagonista REAL de la serie? 🤔

¿Qué relación hay entre las visualizaciones y la valoración? 📺

¿Qué valoraciones tiene la serie? ¿Qué personaje es el mejor valorado? ⬆️

¿Quiénes protagonizan más cantidad de capítulos? 🎭

¿Cuál es la evolución de cada personaje a lo largo de la serie? 🌱

¿Qué presencia tienen el resto de personajes en la serie? 🙌



## ¿Qué presencia tienen el resto de personajes en la serie? 🙌

Empezamos por el final...

Dejando un momento a un lado a los 6 protas, comenzamos hablando del resto de personajes que participan en la serie. De ellos hemos estudiado el porcentaje de apariciones que tienen.

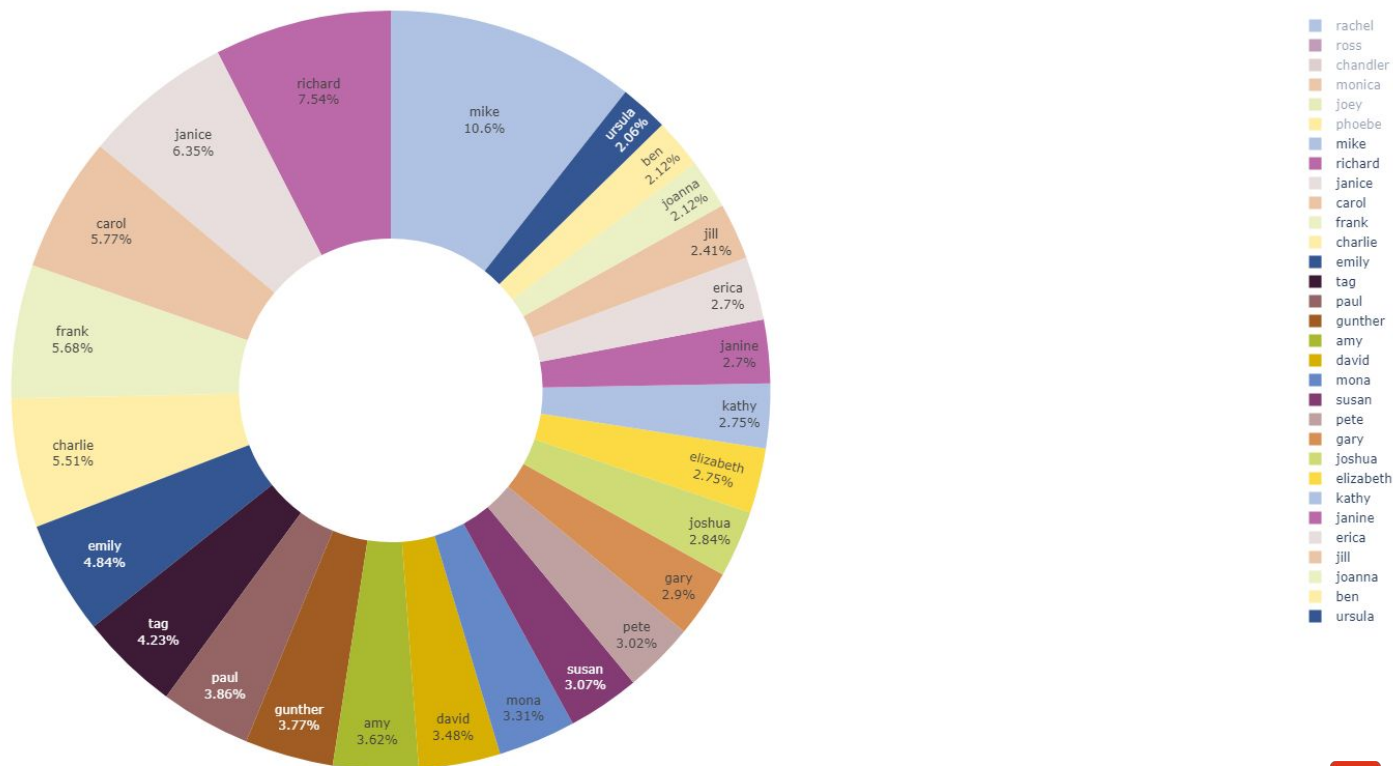
Se han tenido en cuenta solo los personajes que hubieran tenido como mínimo 70 intervenciones a lo largo de todas las temporadas, ya que los datos del resto de personajes eran demasiado bajos para tenerlos en cuenta.





# Resultados

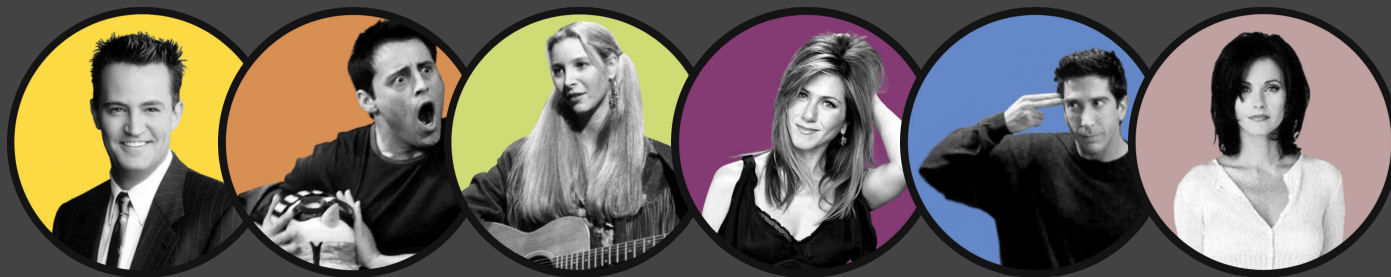
Recuento y porcentaje de aparición del resto de personajes



## ¿Cuál es la evolución de cada personaje a lo largo de la serie? 🌿

Mediante nuestro primer dataset y posteriormente al limpiado completo, hemos construido una nueva tabla de datos con los 6 personajes principales, a saber, estos de aquí ➡

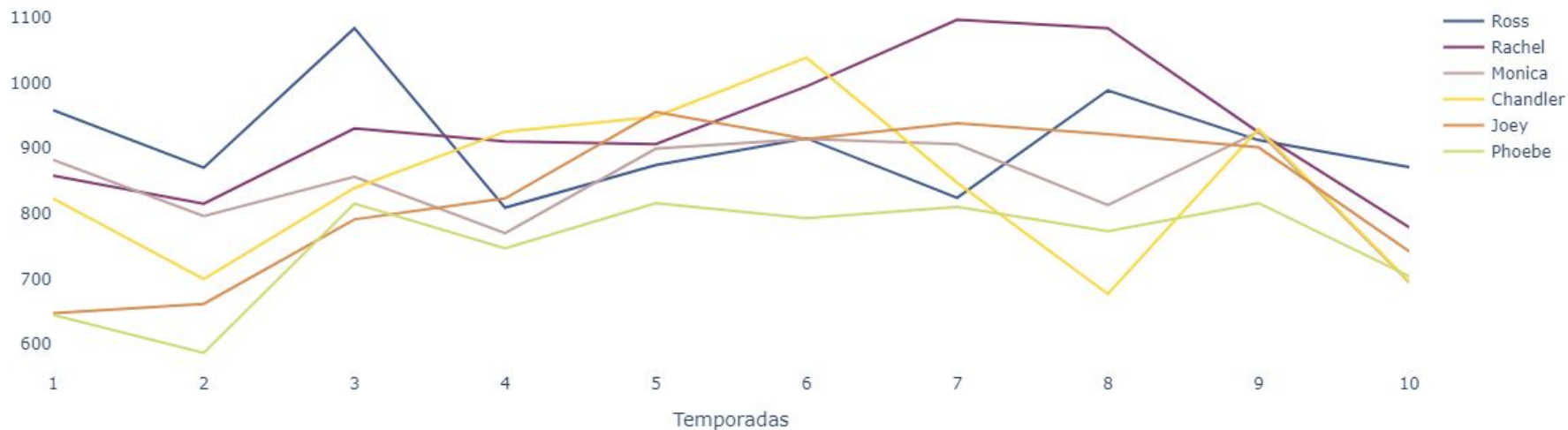
También hemos sacado los datos de los protagonistas de cada capítulo (aquellos que más apariciones tienen en él) y hemos obtenido un gráfico con esos datos.



# Resultados

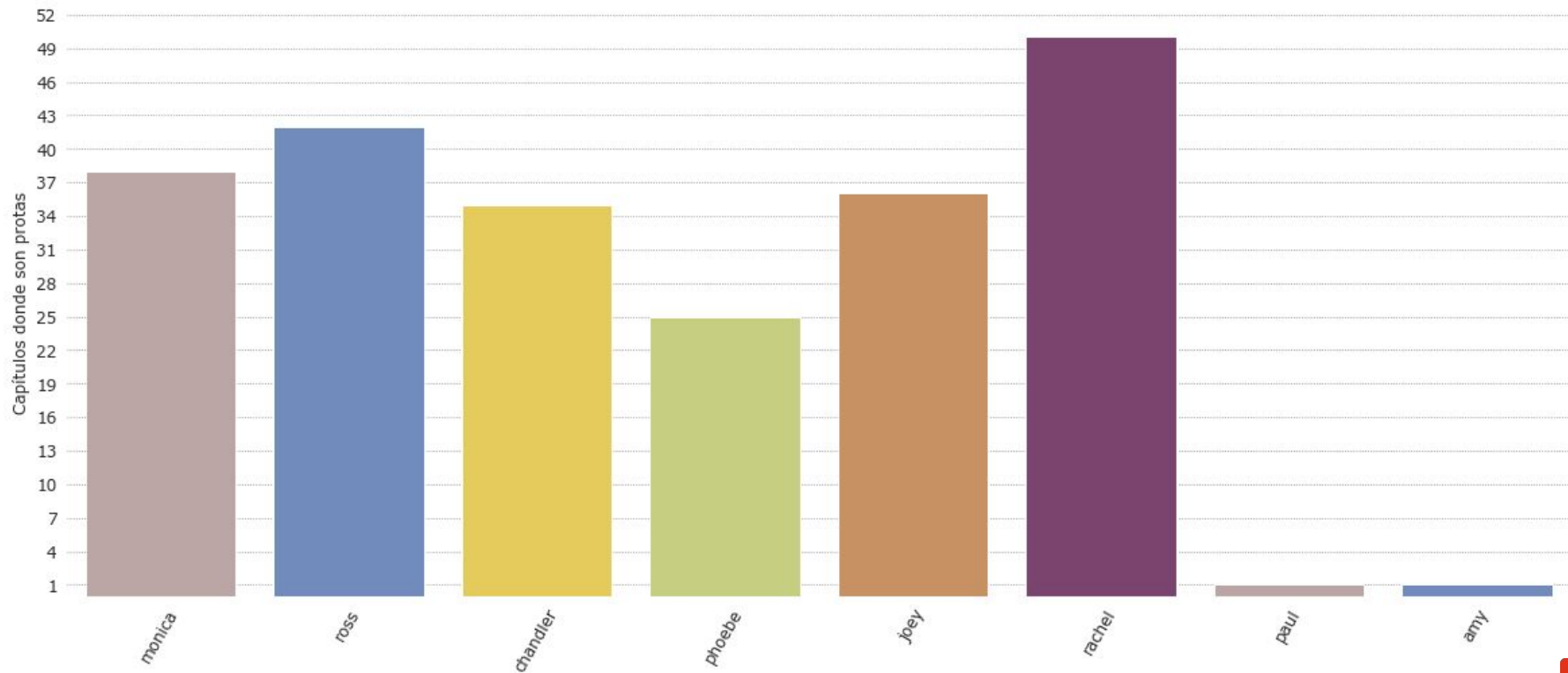
El gráfico se ha obtenido a través del conteo de apariciones por cada capítulo y temporada, de cada uno de los 6 protas.

Aparición de cada personaje por temporada



## Resultados

Quién protagoniza más o menos capítulos a lo largo de toda la serie.



## ¿Qué personaje es el más valorado? ↑

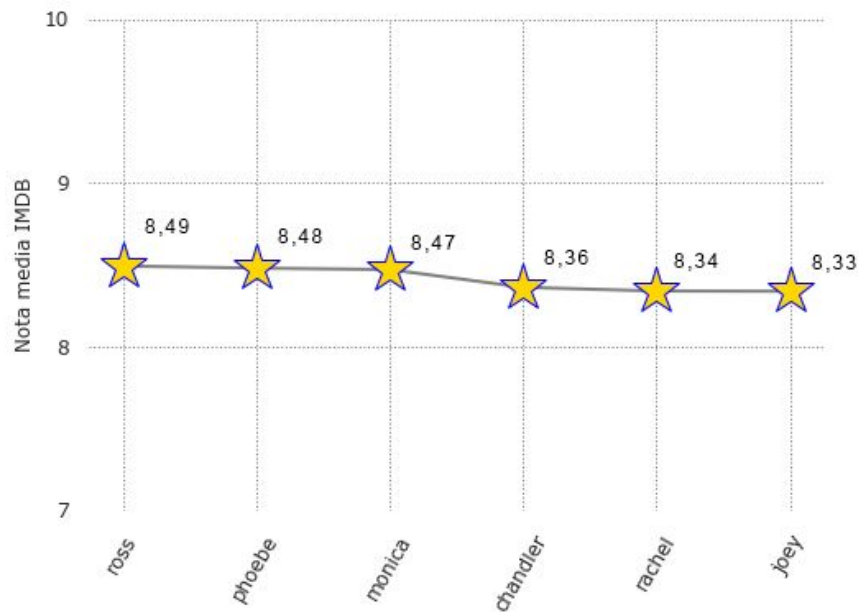
Mediante la unión del dataset de valoraciones y el dataset inicial con los guiones, hemos asociado las valoraciones de cada capítulo, con el protagonista del capítulo.

Es una hipótesis en la que planteamos que la valoración del capítulo está relacionada con el prota del mismo, y por lo tanto, de esta manera, obtenemos las valoraciones medias por personaje, basándonos en los capítulos que protagonizan.

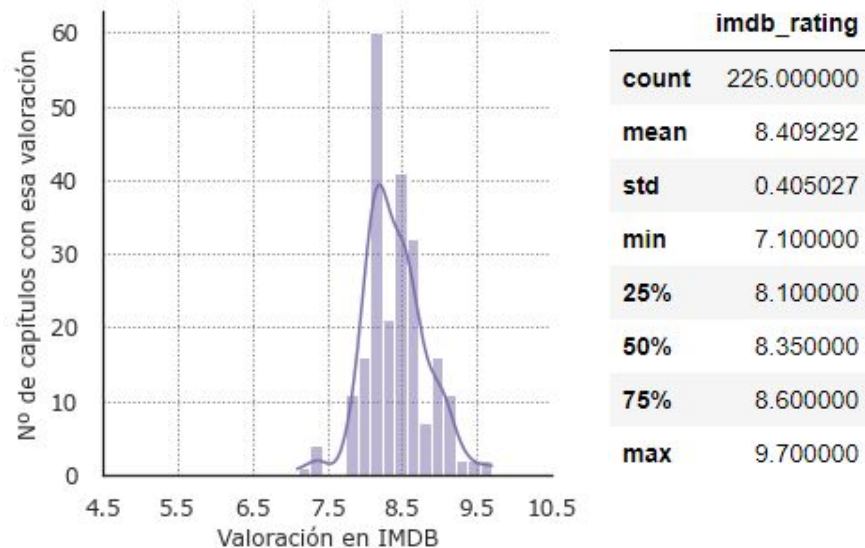


## Resultados

Valoración media de cada prota en IMDB



Total de valoraciones por cap. en IMBD y datos estadísticos



## ¿Qué relación hay entre las visualizaciones y la valoración? 📺

Se han estudiado igualmente los datos de visualizaciones totales por temporadas.

La gráfica que veréis a continuación, nos muestra la media de visualizaciones que tuvo cada una de las temporadas

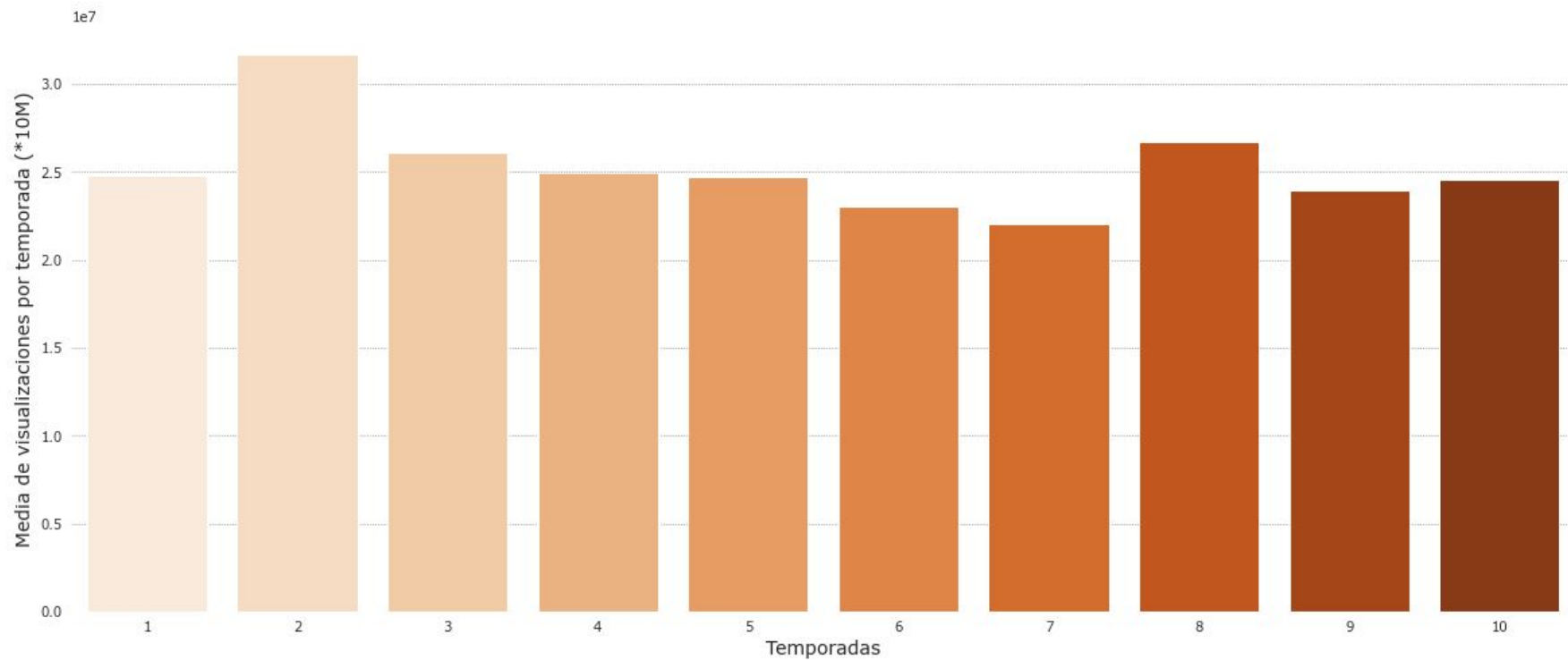
También, unificando nuestro dataset actual con el último dataset que contiene los datos de visualizaciones, podemos obtener las relaciones entre las valoraciones en IMDB y los millones de visualizaciones.

Se puede observar que existe cierta correlación positiva entre el número de visualizaciones y el incremento de la nota de valoración en IMBD. También podemos ver que a la derecha del gráfico hay una serie de valores que están muy alejados de la media.



# Resultados

Media de visualizaciones por temporada

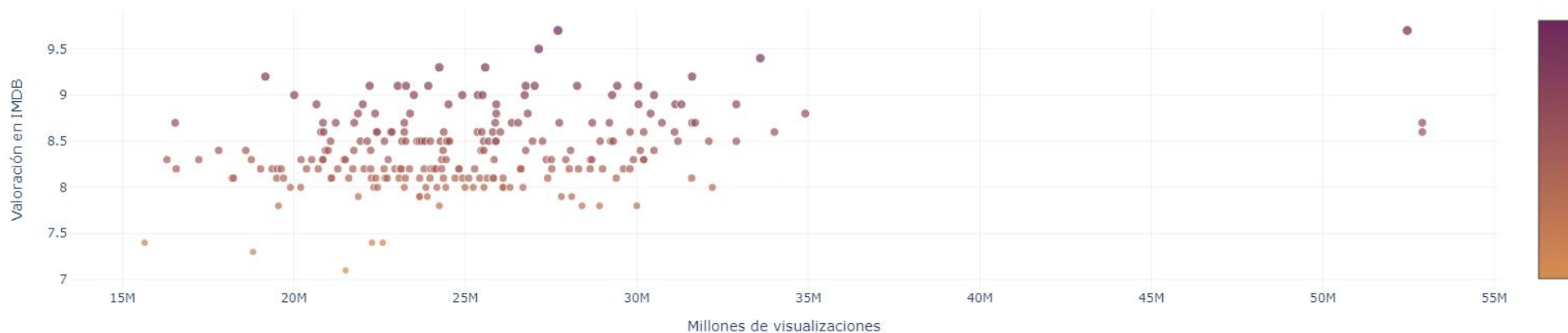




# Resultados

Correlación entre valoraciones en IMBD y millones de visualizaciones

Relación entre valoración en IMDB y visualizaciones



## ¿Quién es el protagonista REAL de la serie? 🤔

Para finalizar el estudio, responderemos a la hipótesis principal del mismo...  
¿Quién es el auténtico protagonista?

Cuando empezamos a analizar los datos que teníamos para poder responder a esta pregunta, nos dimos cuenta de que había dos posibles respuestas y decidimos estudiar ambas y comparar los resultados.

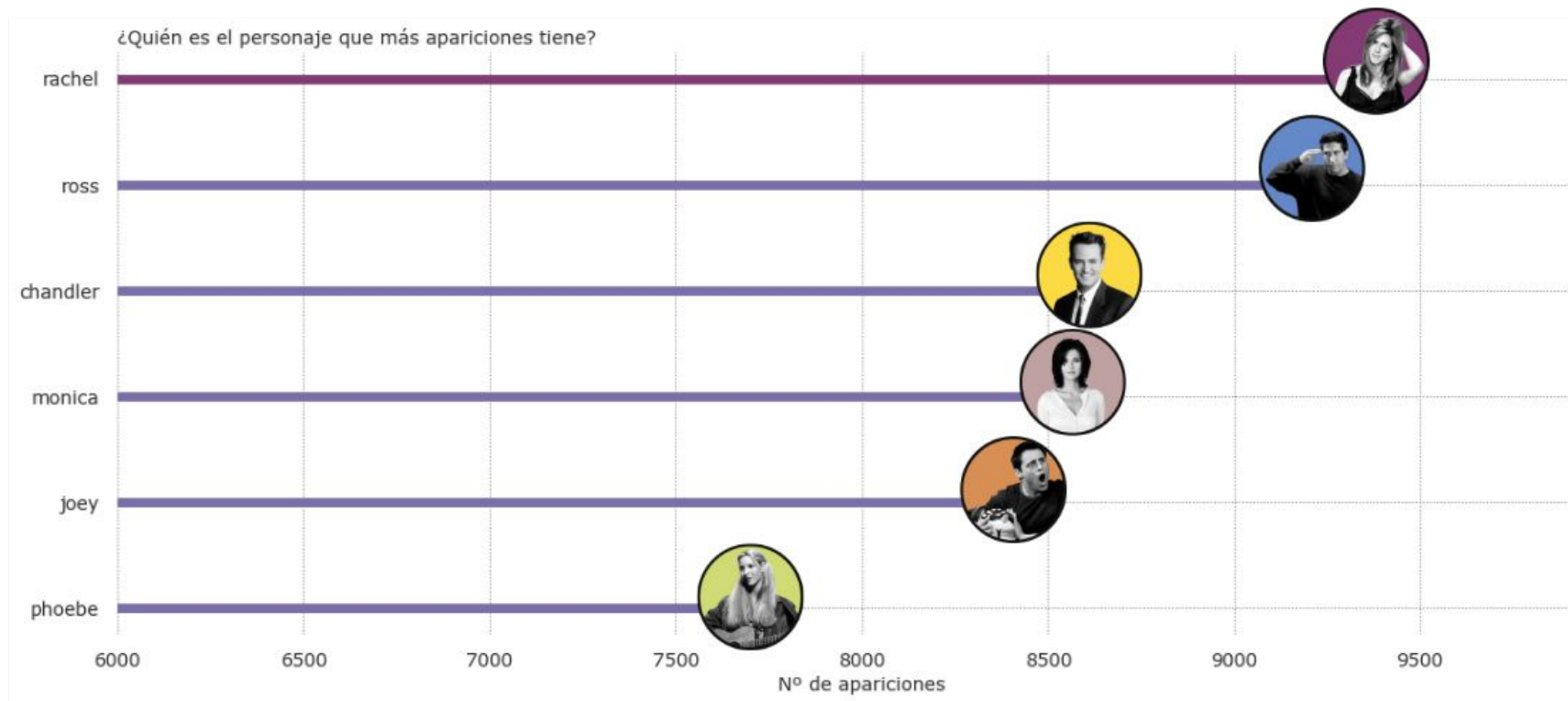
Para ello, hemos hecho dos gráficas muy similares pero con un origen de datos distinto. En la primera de ellas hacemos un conteo de todas las veces que uno de los protas interviene a lo largo de toda la serie (datos que ya hemos tenido en cuenta anteriormente para, por ejemplo, la gráfica de protagonistas de capítulos.

En la segunda, en cambio, hacemos un conteo de la cantidad de palabras en total que dice cada uno ¡a lo largo de todas las temporadas!




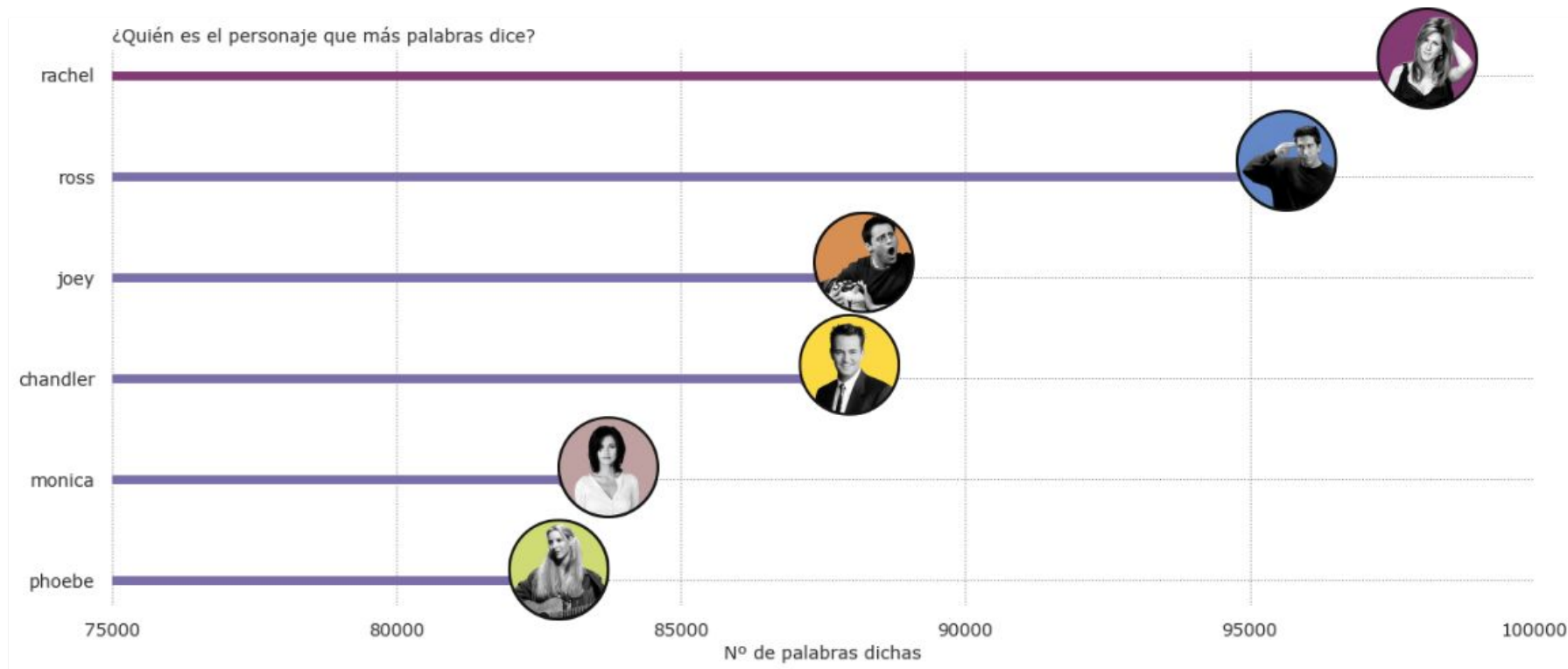
## Resultados

El personaje que más veces interviene en la serie es  con un total de **9295** intervenciones.



## Resultados

El personaje que más palabras dice en la serie es (otra vez)  con un total de **97525** palabras.



## BONUS TRACK 2.0

Para finalizar, os dejo por aquí los datos de la encuesta realizada a la clase... ¿Alguien habrá acertado 🤪?

De un total de 9 respuestas... los resultados han sido:



**Joey** ▶▶ 11% de votos



**Rachel** ▶▶ 0% de votos



**Phoebe** ▶▶ 0% de votos



**Monica** ▶▶ 33% de votos



**Ross** ▶▶ 23% de votos



**Chandler** ▶▶ 33% de votos



Gracias 😊

