

TRABAJO PRÁCTICO INTEGRADOR



Alemán Fernando Ebert - 80753

Arrascaeta Ana Paula - 85523

Indelangelo Nicolas - 55186

GRUPO 13 - 5K3

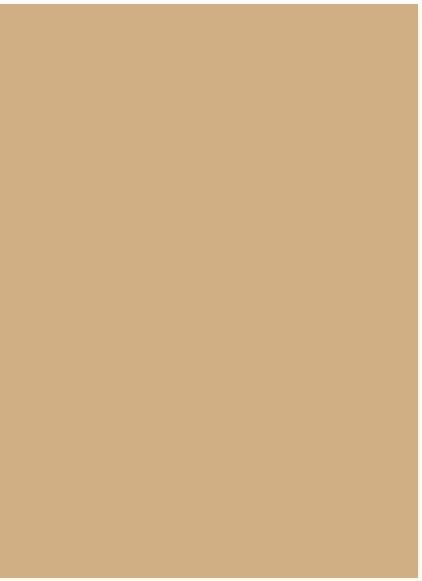
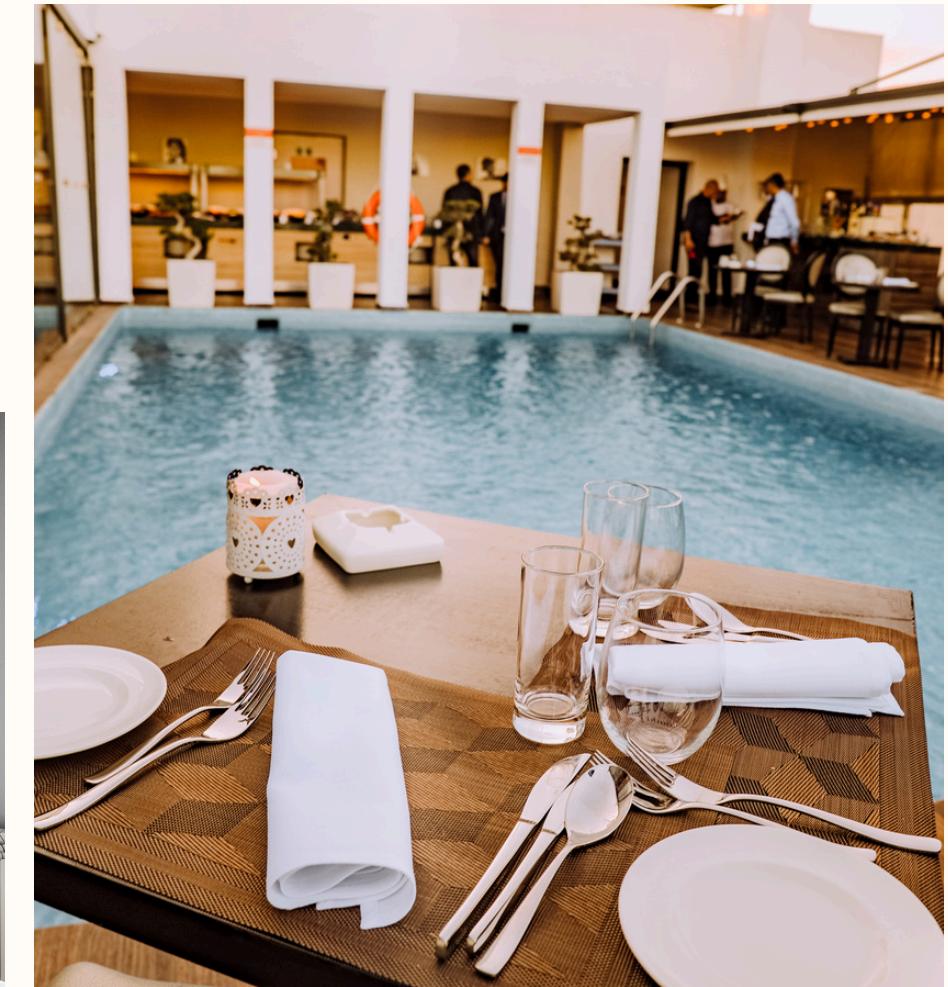
HOTEL REVIEWS PREDICTOR



¿QUÉ HACE QUE UN HUESPED VUELVA A UN HOTEL?

LAS EMOCIONES DETRÁS DE UNA
RESEÑA PUEDEN TENER LA
RESPUESTA

A lo largo de este informe presentaremos cómo el análisis de sentimiento aplicado a reseñas de hoteles puede transformarse en un sistema de recomendaciones personalizadas que no solo predice calificaciones, sino que también anticipa las preferencias de los huéspedes



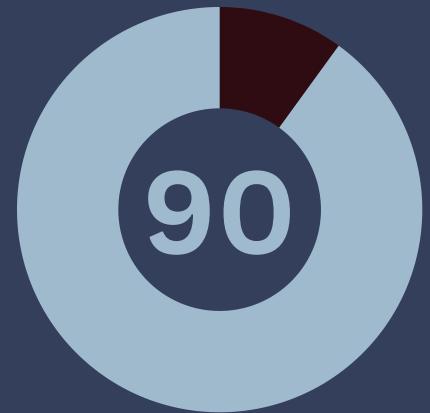
¿POR QUÉ ES IMPORTANTE?

Las reseñas de huéspedes son una mina de oro de información, pero, ¿Cómo podemos ir más allá de que solo sean calificaciones numéricas?"

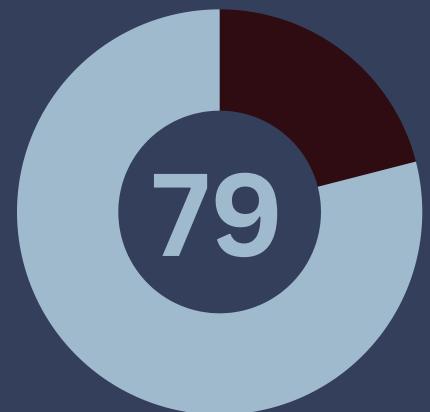


Al combinar el análisis de sentimiento de las reseñas con las características de cada hotel y el historial de interacciones de los huéspedes, podemos construir una experiencia única y personalizada para cada persona."

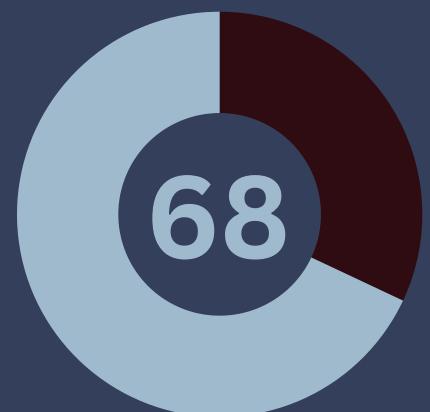




LEEN RESEÑAS EN LÍNEA ANTES DE DECIDIR SOBRE UN HOTEL



CONFÍAN TANTO COMO EN LAS RECOMENDACIONES PERSONALES.



DEJARÍA DE USAR EL SERVICIO TRAS DEJAR MALA RESEÑA.

INTRODUCCIÓN

Actualmente, la información que se crea a diario es de tal magnitud que hay una necesidad extrema de convertir esa magnitud de información en algo significativo que pueda ser aplicado para la toma de decisiones.

La ciencia de datos es una de las disciplinas interdisciplinarias que han emergido en la última década para extraer información significativa de una gran cantidad de datos.

El propósito de este proyecto es desafiarnos a tomar una actitud práctica hacia el tratamiento y análisis de datos, junto con el desarrollo de conocimientos de los procesos, técnicas y algoritmos utilizados en el campo de la ciencia de datos.

Nuestra meta con este trabajo integrador es hacer todo el ciclo de vida de un proyecto de ciencia de datos sobre un conjunto de datos que ya se encuentra definido. Como parte de este proceso ejercitaremos nuestra capacidad para analizar datos desde la recopilación y limpieza hasta el análisis e interpretación con el fin de derivar información útil y no obvia que pueda ayudar a resolver problemas particulares, identificando patrones y tendencias, también nos ayudará a incorporar una nueva visión para la toma de decisiones desde el punto de vista de los datos

DESCRIPCIÓN DEL DATASET

El dataset elegido contiene 515,000 reseñas de clientes y puntuaciones de 1493 hoteles de lujo en Europa, con 17 campos que describen tanto aspectos de los hoteles como de las reseñas proporcionadas por los usuarios. Estos datos fueron obtenidos de Booking.com y contienen información geográfica y textual que puede utilizarse para realizar un análisis exhaustivo sobre la satisfacción del cliente. Fuente: 515K Hotel Reviews Data in Europe (kaggle.com)

DESCRIPCIÓN DE LAS VARIABLES

Hotel_Address:	Dirección del hotel.
Review_Date	Fecha de publicación de la reseña.
Average_Score	Puntuación media del hotel basada en las reseñas del último año. Hotel_Name: Nombre del hotel.
Reviewer_Nationality	Nacionalidad del revisor
Negative_Review	Comentario negativo proporcionado por el usuario
Review_Total_Negative_Word_Counts	Número total de palabras en la reseña negativa.
Positive_Review	Comentario positivo proporcionado por el usuario
Review_Total_Positive_Word_Counts	Número total de palabras en la reseña positiva
Reviewer_Score	Puntuación que el revisor ha dado al hotel
Total_Number_of_Reviews_Reviewer_Has_Given	Número total de reseñas que ha dado el revisor en el pasado
Total_Number_of_Reviews	Número total de reseñas válidas del hotel
Tags	Etiquetas que el revisor asignó al hotel
Days_Since_Review	Duración entre la fecha de la reseña y la fecha de recolección del dato
Additional_Number_of_Scoring	Número de puntuaciones válidas sin comentarios
Lat	Latitud del hotel
Lng	Longitud del hotel

ANÁLISIS INICIAL

Con nuestro conjunto de datos, podemos lograr una variedad de objetivos y obtener insights valiosos sobre la satisfacción del cliente en hoteles de lujo en Europa, a continuación en detalle.

SEGMENTACIÓN DE CLIENTES

El objetivo es segmentar a los clientes en base a variables como nacionalidad, frecuencia de reseñas, o tipo de comentarios. Esto nos permite personalizar la oferta y los servicios del hotel para satisfacer mejor las necesidades de diferentes grupos de clientes.

OPTIMIZACIÓN DE ESTRATEGIAS DE MARKETING

El objetivo es analizar la efectividad de las estrategias de marketing y promociones en función de las reseñas y calificaciones. Esto permite ajustar las estrategias de marketing basándose en la retroalimentación real de los clientes, mejorando así el retorno de inversión en campañas promocionales

MODELO PREDICTIVO DE CALIFICACIÓN

El objetivo es desarrollar un modelo para predecir la calificación de una reseña basada en los datos históricos de reseñas y calificaciones. Como beneficio, Booking puede decidir mejor que ofrecer a sus clientes, y un beneficio extra es que los hoteles pueden anticipar la calificación de nuevas reseñas, lo que puede ayudar a gestionar la reputación y mejorar las áreas que impactan negativamente en la satisfacción del cliente.

IDENTIFICACIÓN DE FACTORES CLAVE DE SATISFACCIÓN

El objetivo es analizar qué variables tienen mayor impacto en la puntuación dada por los usuarios. Esto puede incluir aspectos como comentarios positivos/negativos, nacionalidad del revisor, o número de palabras o su utilización en las reseñas. Esto permite a los hoteles entender qué factores influyen más en la satisfacción del cliente y enfocar sus esfuerzos de mejora en áreas clave

GESTIÓN DE PROYECTO

MARCO DE GESTIÓN

Para la gestión del proyecto, implementaremos un enfoque Scrum Data Driven Agile, que es una variante de un marco Scrum ágil, incorporando un enfoque basado en datos y la recopilación y análisis de estos para ajustar el proceso de desarrollo. El Scrum Data Driven Agile también pone énfasis en la mejora continua a través de la medición de resultados y el cambio de estrategia según lo dictado por los datos. En cada sprint, se analizan datos para encontrar problemas, verificar hipótesis y mejorar el proceso de desarrollo basándose en hechos.



¿PORQUÉ LO UTILIZAMOS?

PRIORIZACIÓN DINÁMICA

Las prioridades pueden cambiar basándose en los descubrimientos realizados durante el análisis de datos. Esta técnica permite la reprogramación dinámica del backlog y la priorización inmediata de las tareas más críticas, asegurando la optimización de recursos y tiempo

CAPACIDAD DE ADAPTACIÓN

La ciencia de datos requiere una gran adaptabilidad, ya que los resultados del análisis pueden sugerir un cambio en el enfoque del proyecto. El Scrum permite una rápida adaptación, y los datos garantizan que cualquier cambio esté respaldado por la información más actual

ENFOQUE BASADO EN PRUEBAS

En ciencia de datos, las decisiones deben tomarse en función de los análisis recopilados a lo largo del proyecto. El Scrum basado en datos permite flexibilidad en los objetivos e iteraciones según los resultados de experimentos y modelos, asegurando que las decisiones se tomen en función de datos reales.

HERRAMIENTAS QUE UTILIZAMOS

PYTHON

Utilizaremos Python con distintas librerías como Pandas para procesar datos, NumPy para cálculos, Matplotlib y Seaborn para visualización, y Scikit-learn para modelado predictivo.

GOOGLE DRIVE

Google Drive se usará para almacenar y compartir documentos e informes del proyecto.

JUPYTER NOTEBOOKS

Usaremos Jupyter Notebooks para documentar, visualizar y experimentar con el desarrollo de modelos, facilitando la organización del código y comentarios.

CANVA

Canva se utilizará para crear imágenes, documentos y presentaciones que nos ayuden a representar el avance del proyecto y sus resultados

GOOGLE COLAB

Nos permite ejecutar código en la nube y colaborar en tiempo real, ideal para análisis de datos y modelado predictivo con Python.

>>> STAKEHOLDERS

EQUIPO DE
TRABAJO

CONSULTORES DE
HOSPEDAJE EN LÍNEA

USUARIOS DE
KAGGLE

PROFESORES
INTERESADOS

CREADORES DEL
DATASET

ESTUDIANTES

COMUNIDAD DE
ANÁLISIS DE DATOS

CLIENTES DE
HOTELERÍA

HOTELES DE EUROPA

REVISORES DE LA
INDUSTRIA
HOTELERA

INICIO DEL PROYECTO

SPRINT 1

EXPLORACIÓN Y PREPARACIÓN DE DATOS

US: Como científico de datos, necesito generar características útiles para mejorar el rendimiento del modelo y procesar los datos adecuadamente.

TAREAS

1	Realizar una revisión inicial del Dataset, en búsqueda de valores inconsistentes y nulos
2	Limpieza de datos (eliminar duplicados, tratar datos faltantes, conteo de blancos en las reseñas).
3	Análisis exploratorio de datos para identificar patrones iniciales (EDA).
4	Realizar resumen estadístico de las variables más importantes (medias, medianas, distribuciones, etc)
5	Generar visualizaciones básicas de las variables más importantes (score promedio del hotel, distribuciones de reseñas por extensión, score otorgado por el revisor, cantidad de revisiones del revisor).
6	Definir cuáles variables a utilizar en el modelo predictivo.

TAREAS

SPRINT 2

FEATURE ENGINEERING Y PRE PROCESAMIENTO

US: Como científico de datos, necesito preparar y explorar el dataset para asegurar que los datos sean de calidad y estén listos para el análisis

1	Transformar las variables categóricas en numéricas (OneHot Encoding o Label Encoding).
2	Reescalar variables numéricas (Normalización).
3	Crear nuevas características a partir de las reseñas
4	Analizar la correlación entre las variables (usando mapas de calor o gráficos de dispersión).
5	Dividir el Dataset en conjuntos de entrenamiento y prueba
6	Realizar validación cruzada para asegurar la consistencia de los datosvariables a utilizar en el modelo predictivo.

TAREAS

SPRINT 3

SELECCIÓN Y ENTRENAMIENTO DEL MODELO

US: Como científico de datos, necesito seleccionar y entrenar un modelo predictivo con los datos para obtener resultados precisos.

1	Selección del modelo de regresión (regresión lineal, bosques aleatorios, aumento de gradiente)
2	Entrenamiento del modelo con el conjunto de entrenamiento.
3	Ajustar los hiper parámetros.
4	Validar el modelo con los datos de prueba
5	Medir el rendimiento del modelo utilizando métricas (a definir)
6	Hacer iteraciones con diferentes modelos en caso que el rendimiento no sea el adecuado.

TAREAS

SPRINT 4

OPTIMIZACIÓN DEL MODELO Y VISUALIZACIÓN DE RESULTADOS

**US: Como científico de datos,
necesito optimizar el modelo y
visualizar los resultados para
presentar insights útiles**

1	Optimización de los hiper parámetros para mejorar el rendimiento del modelo.
2	Evaluar si es necesario ajustar el modelo o cambiar de enfoque.
3	Crear informes visuales de las predicciones y comparaciones con los valores reales.
4	Visualización de los insights clave a partir del análisis (gráficos, tablas, etc.).
5	Preparar el informe de resultados con conclusiones sobre el comportamiento y la predicción del dataset

TAREAS

SPRINT 5

IMPLEMENTACIÓN FINAL Y PRESENTACIÓN

**US: Como científico de datos,
necesito implementar la
solución final y presentar los
resultados de forma clara y
precisa**

1	Refinar los resultados obtenidos para la presentación final
2	Terminar documentación del proceso seguido en el proyecto (limpieza, preprocesamiento y modelado).
3	Refinar el informe final con todas las visualizaciones, análisis y predicciones realizadas.
4	Preparar la presentación de los resultados más importantes y las conclusiones finales del proyecto.

PREPARACION Y LIMPIEZA

Eliminación de columnas innecesarias

1

Durante el primer análisis se eliminaron:

Additional_Number_of_Scoring, Hotel_Address, lat, lng, days_since_review

Datos faltantes y valores nulos

2

No se encontraron registros con valores faltantes ni valores nulos durante el análisis.

Eliminación de duplicados

3

526 filas eran duplicaciones exactas de otras filas por lo cual fueron eliminadas.

EXPLORACIÓN INICIAL

PROPIEDADES PRINCIPALES



REVIEWER_SCORE	AVERAGE_SCORE	POSITIVE REVIEW	NEGATIVE REVIEW
mean 8.395502 std 1.637484 min 2.500000 25% 7.500000 50% 8.800000 75% 9.600000 max 10.000000	mean 8.397754 std 0.547961 min 5.200000 25% 8.100000 50% 8.400000 75% 8.800000 max 9.800000	mean 8.395502 std 1.637484 min 2.500000 25% 7.500000 50% 8.800000 75% 9.600000 max 10.000000	mean 8.395502 std 1.637484 min 2.500000 25% 7.500000 50% 8.800000 75% 9.600000 max 10.000000

★ Será nuestra variable a predecir. El promedio (8.4) muestra una tendencia general positiva en las reseñas. Debemos tener en consideración que factores influyen en la dispersión de las puntuaciones ya que la desviación estándar de (1.637) es algo alta. La mediana (8.8) es mayor a la media lo que sugiere una ligera asimetría hacia la derecha.

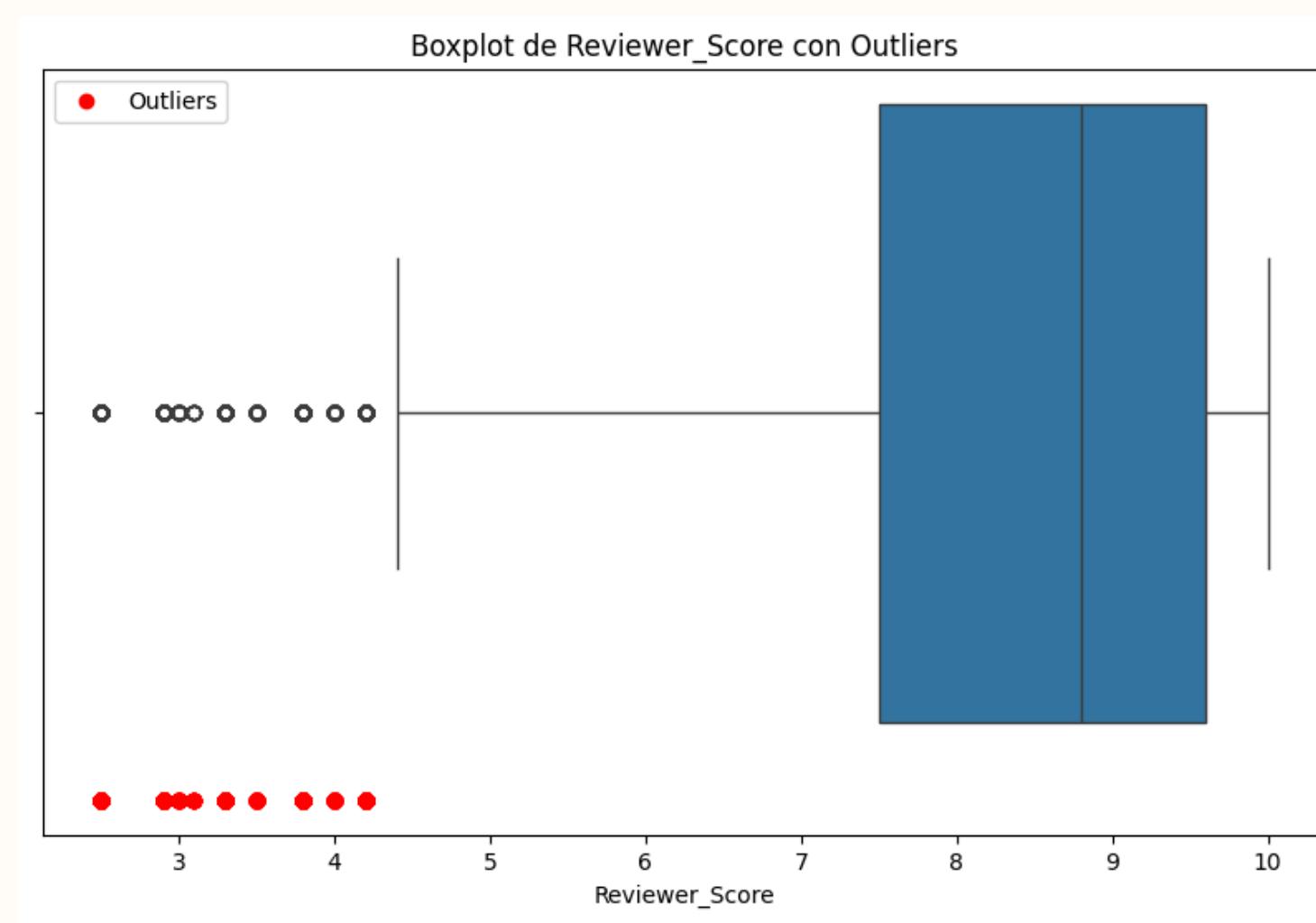
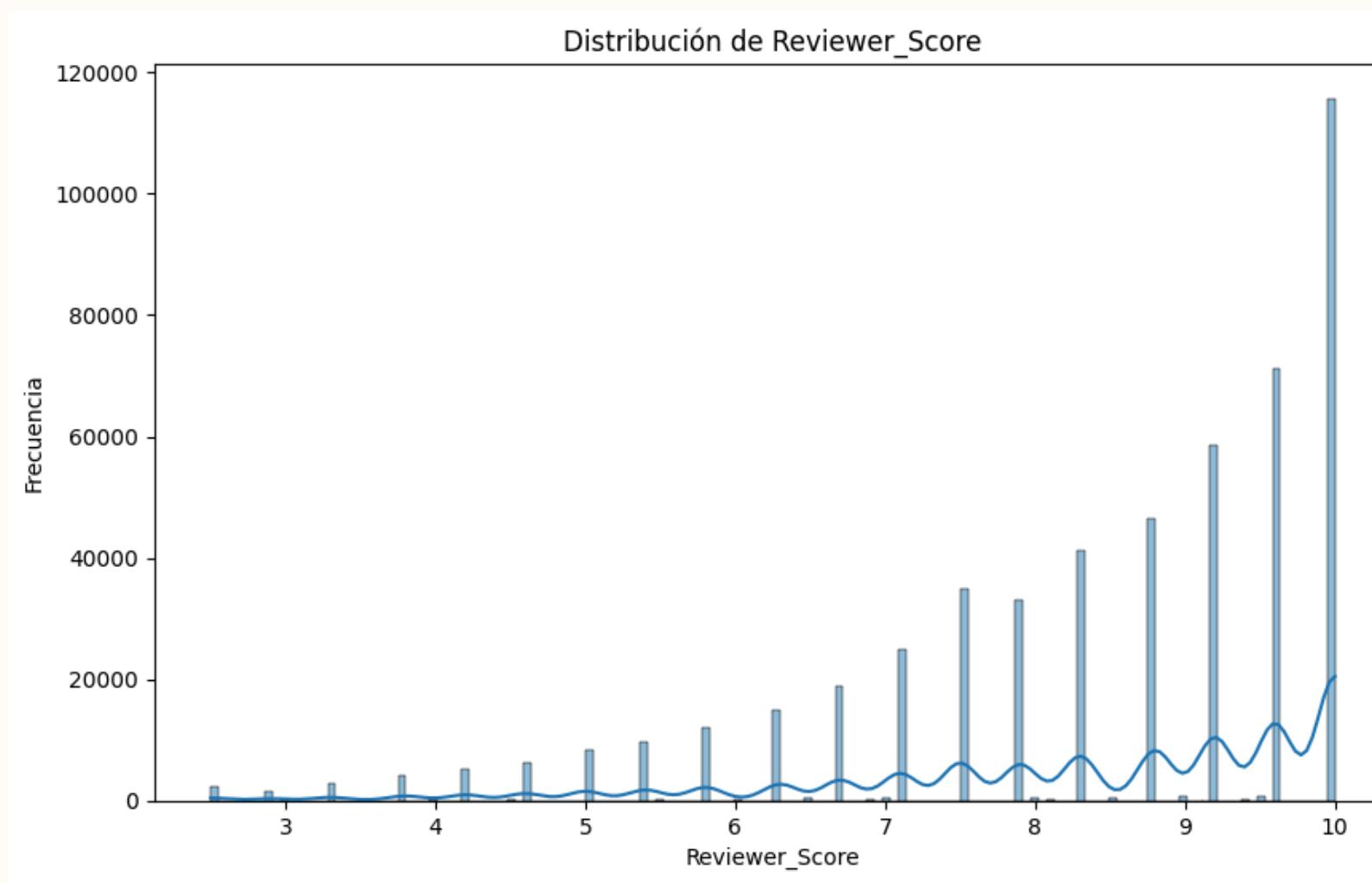
Es un indicador clave de la calidad y reputación de un hotel, útil para predecir la satisfacción del cliente. Su distribución con una media alta (8.4) y una baja desviación estándar (0.547) indica que la mayoría de los hoteles tienen una buena valoración general, lo cual es un factor a considerar en el entrenamiento.

Es crucial para lograr los objetivos del proyecto, ofrece información valiosa sobre los aspectos que los clientes aprecian en cada hotel. El análisis del texto de las reseñas puede revelar patrones y tendencias en la satisfacción del cliente. A partir de ella podremos crear nuevas características e identificar las palabras clave que tienen una mayor influencia positiva sobre las valoraciones.

Es clave para identificar los aspectos que generan insatisfacción en los clientes. El análisis del texto puede revelar patrones y tendencias negativas, permitiendo crear nuevas características, captar el sentimiento e identificar palabras clave con mayor influencia en las valoraciones bajas.

EXPLORACIÓN INICIAL

ANÁLISIS DE DATOS



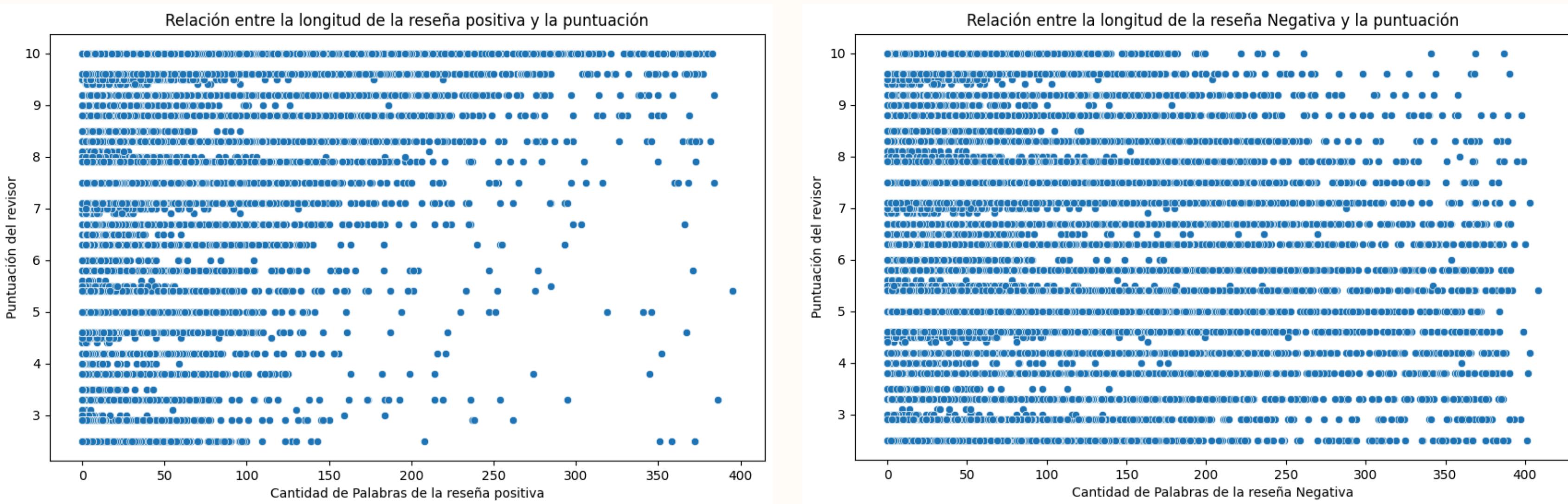
La mayoría de las puntuaciones están concentradas en el extremo superior de la escala, lo que indica una distribución sesgada hacia las evaluaciones positivas lo que puede afectar la capacidad del modelo para generalizar adecuadamente en casos menos comunes, como puntuaciones intermedias o bajas.

Sera necesario balancear el conjunto de entrenamiento o la selección de un modelo que pueda lidiar con este tipo de desbalanceo en los datos.

El boxplot muestra que la mayoría de las calificaciones están por encima de 8, con algunos outliers menores a 4. Estos valores atípicos pueden indicar insatisfacciones clave o ruido en los datos, por lo que es importante analizarlos cuidadosamente para mejorar la precisión del modelo.

EXPLORACIÓN INICIAL

ANÁLISIS DE DATOS

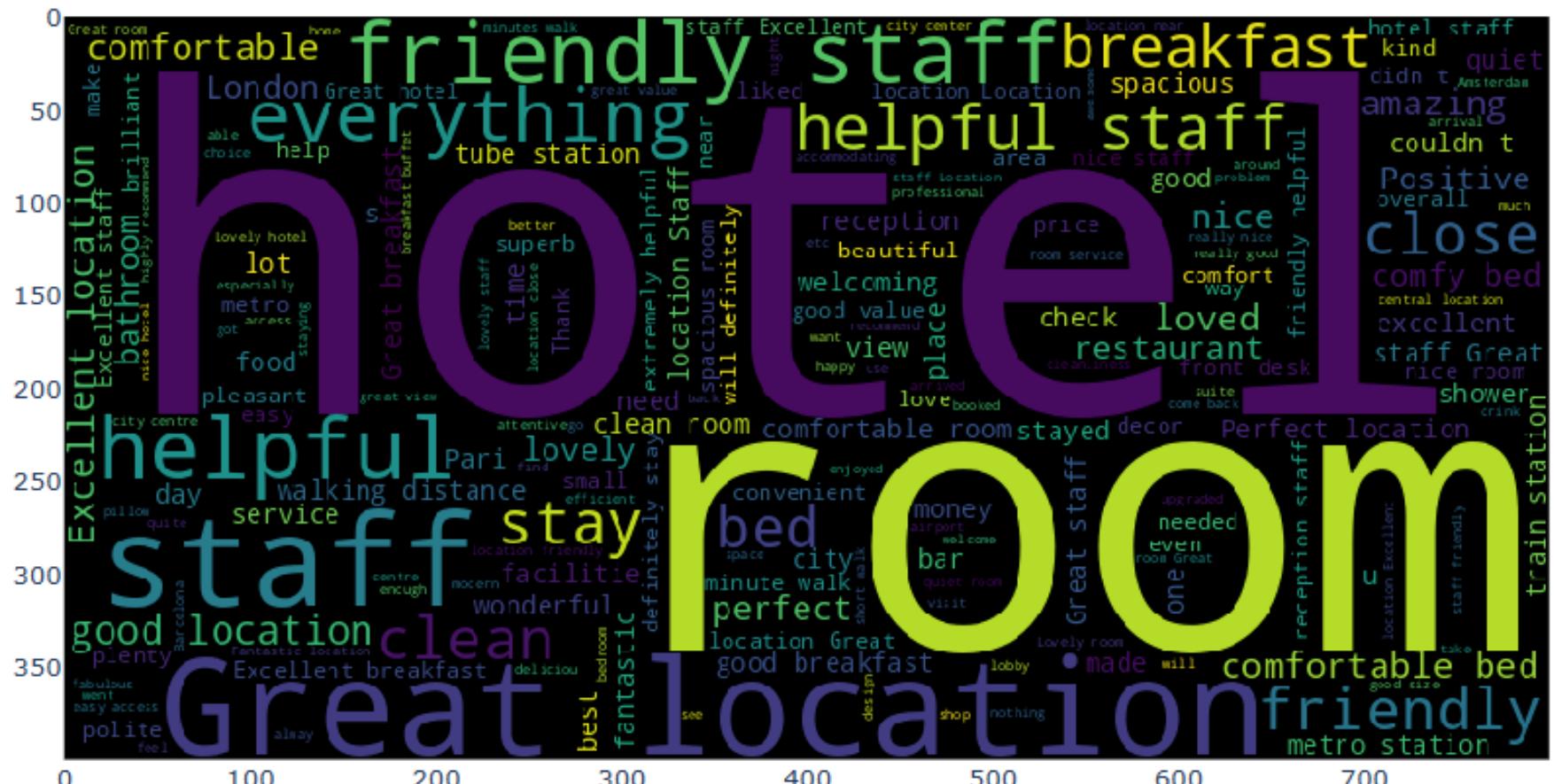


No se aprecia una relación claramente definida entre la longitud de las reseñas (positivas o negativas) y la puntuación otorgada por el revisor. Las distribuciones de puntos parecen estar dispersas en todos los niveles de puntuación, sin un patrón consistente que indique una correlación directa. En el caso de las reseñas negativas, se observa que tanto reseñas cortas como largas se asocian con puntajes variados. De manera similar, para las reseñas positivas, las puntuaciones altas y bajas están presentes en diferentes longitudes de texto. Esto sugiere que la cantidad de palabras en las reseñas no es un factor determinante de la calificación del revisor. Por lo cual definimos que el enfoque correcto será tratar de captar el sentimiento de los huéspedes y buscar patrones desde el texto de las reseñas.

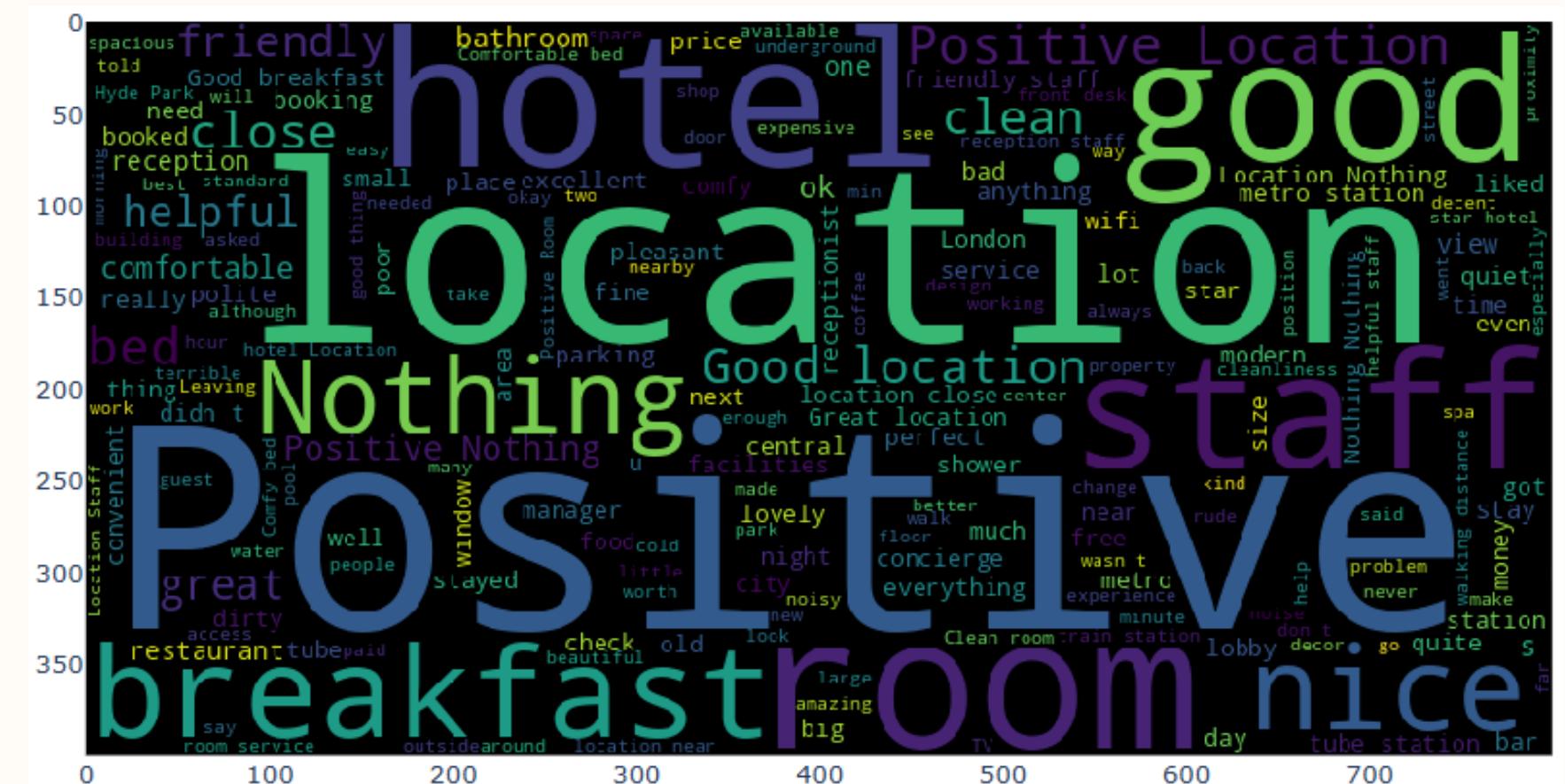
ANÁLISIS DE PALABRAS

NUBE DE PALABRAS

Palabras Positivas en buenas reviews.



Palabras Positivas en malas reviews



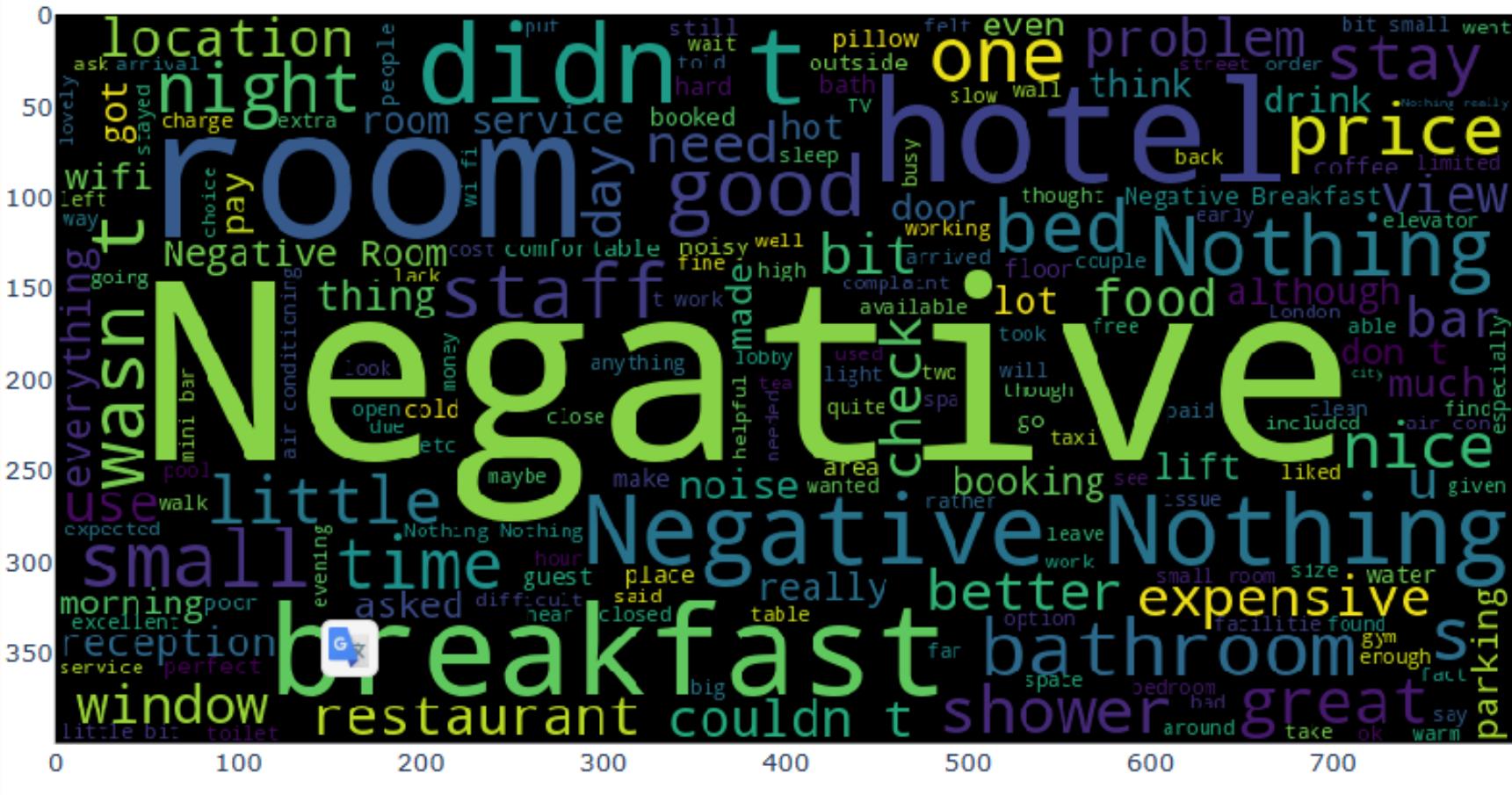
Utilizamos WordClouds para visualizar las palabras más frecuentes, para mejorar la utilidad de los datos categorizamos las reseñas de la siguiente manera:

- Buenas reviews: Review_Score > 8
- Malas reviews (para hoteles de lujo): Review_Score <= 8

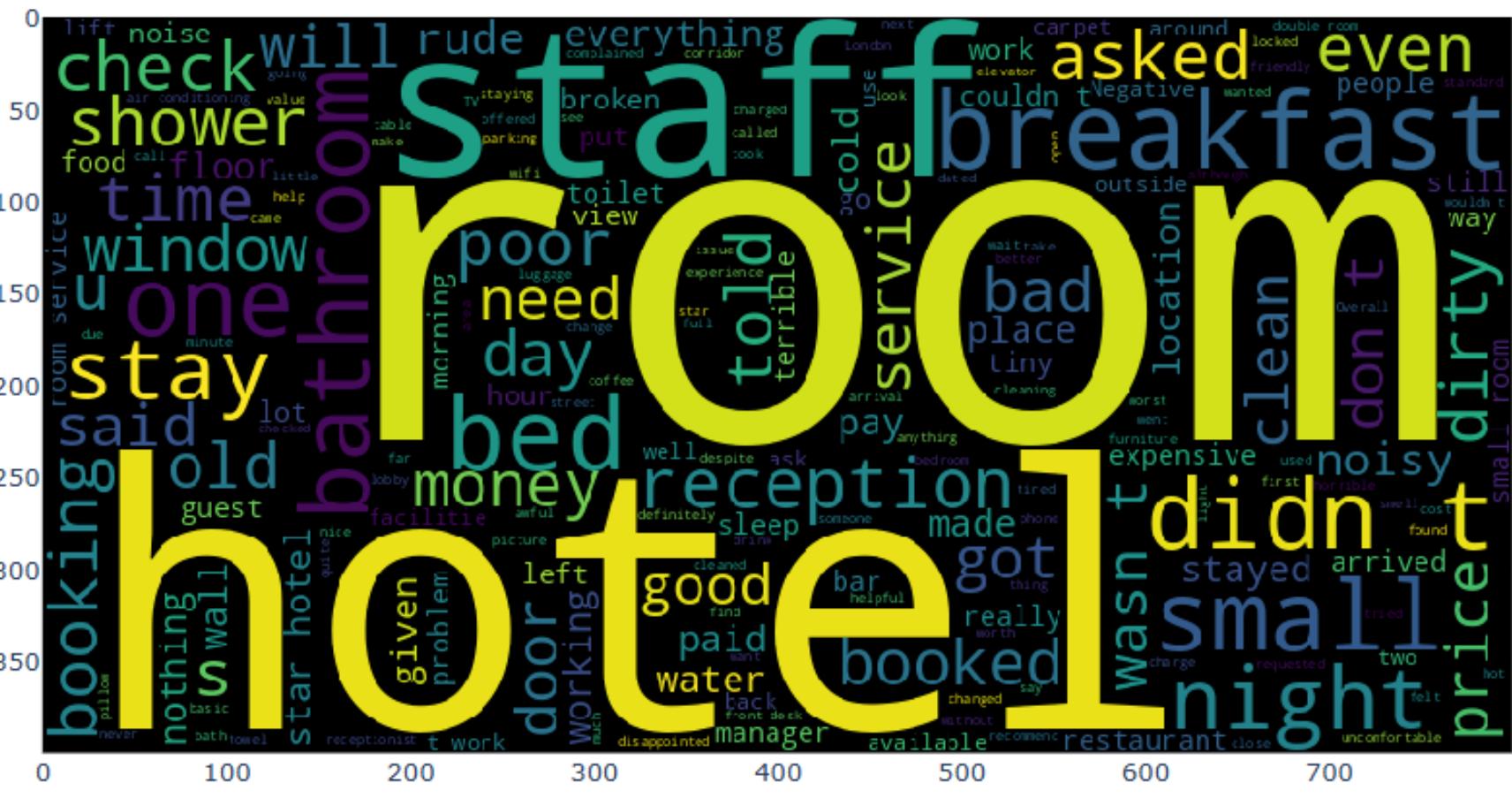
ANÁLISIS DE PALABRAS

NUBE DE PALABRAS

Palabras Negativas en buenas reviews



Palabras Negativas en malas reviews



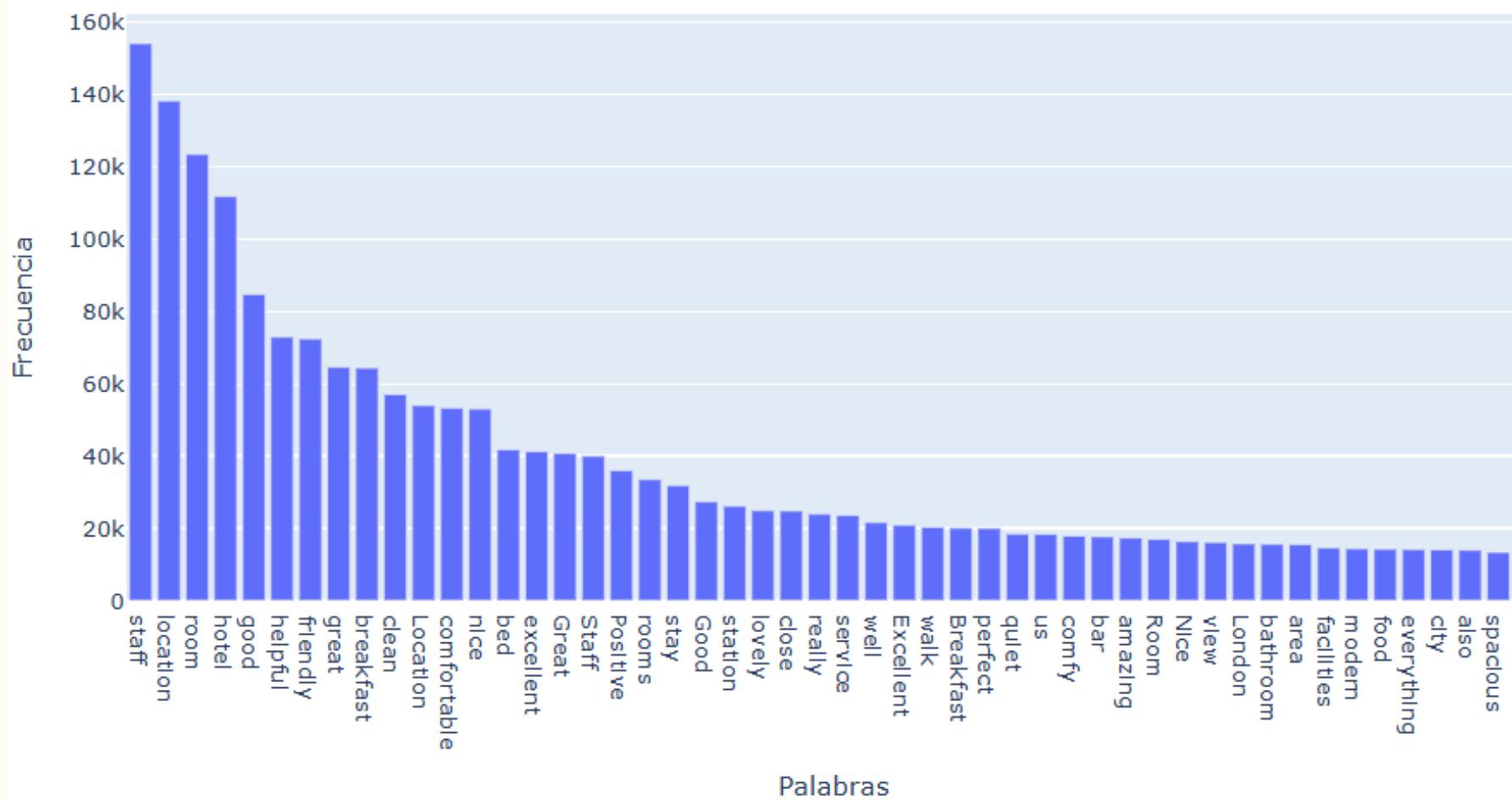
Los gráficos de nube de palabras muestran que, tanto en reseñas positivas como en negativas, ciertas palabras clave aparecen recurrentemente, como "hotel", "room", "location" y "staff". Esto sugiere que los aspectos más comentados por los usuarios se relacionan con elementos básicos de la experiencia, independientemente del tono general de la reseña.

Entonces concluimos que: tanto las reseñas buenas como malas parecen centrarse en aspectos concretos del servicio o las instalaciones, pero el contexto y las combinaciones de palabras determinan si la percepción es favorable o desfavorable.

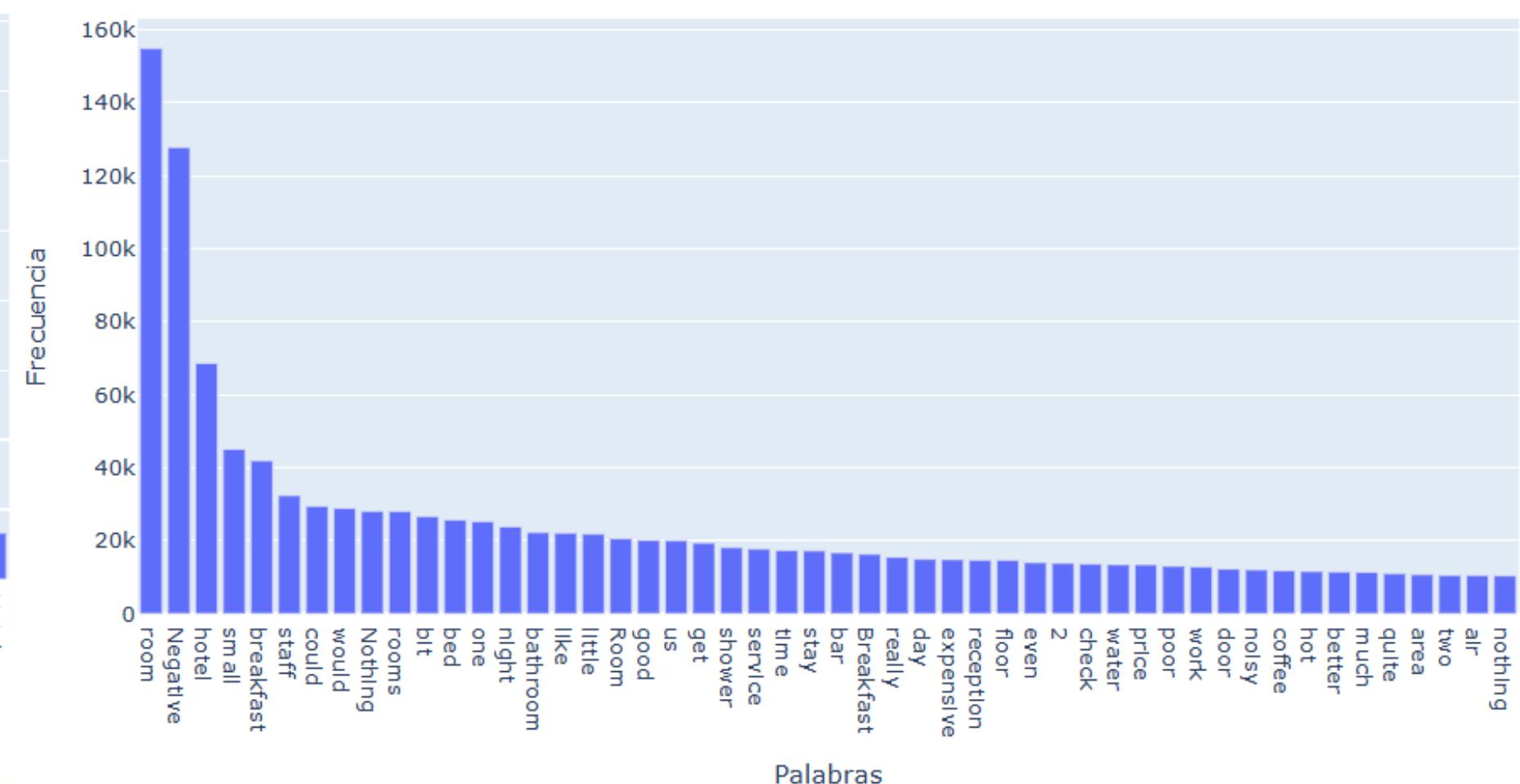
ANÁLISIS DE SENTIMENTO

ANÁLISIS DE FRECUENCIAS

Frecuencia de las 50 palabras más usadas en Positive_Review



Frecuencia de las 50 palabras más usadas en Negative Review

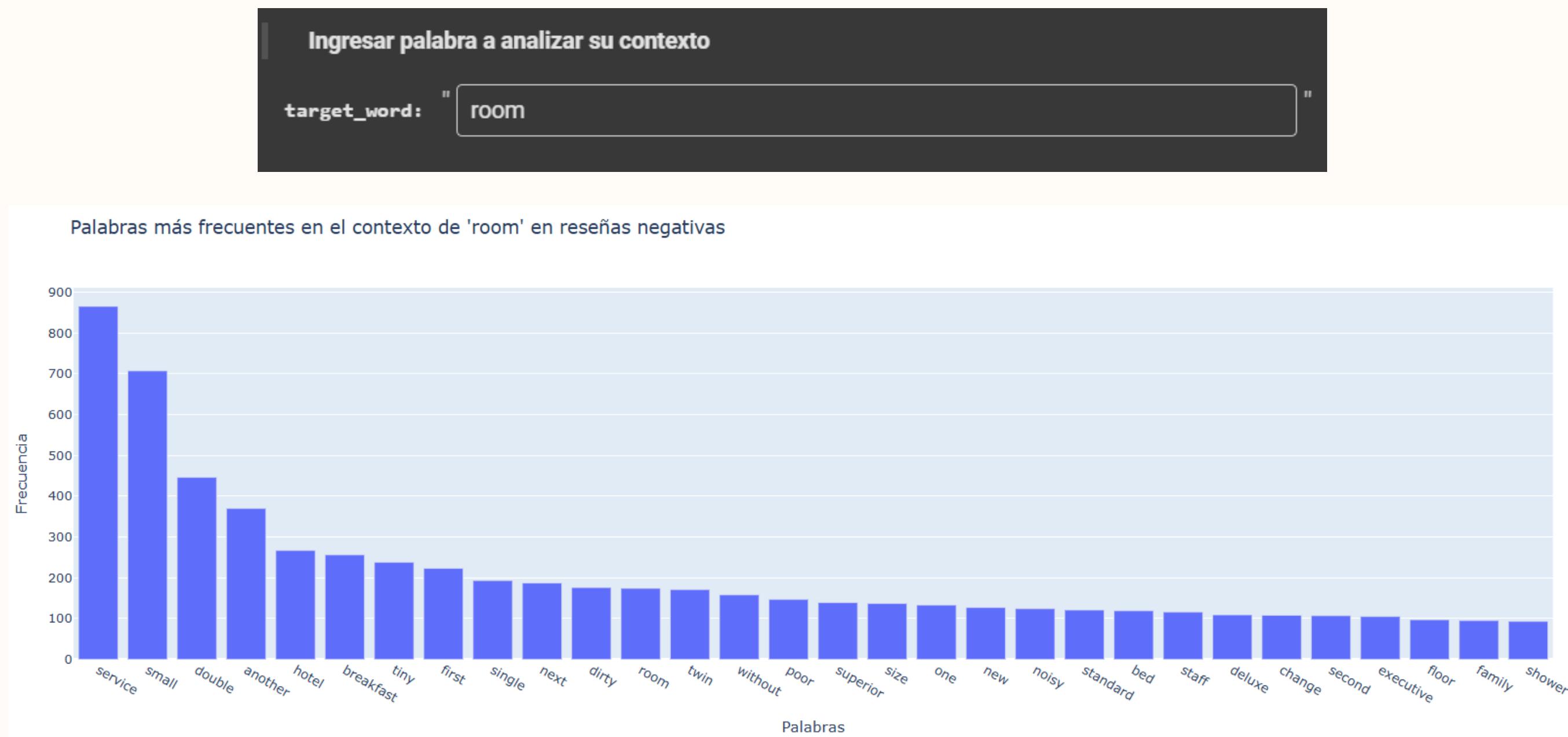


El análisis de la frecuencia de aparición de las palabras nos ayudara a elegir aquellas palabras sobre las cuales realizar un análisis contextual para determinar el sentido de su aparición, tanto en reviews positivas como negativas.

ANÁLISIS DE SENTIMIENTO

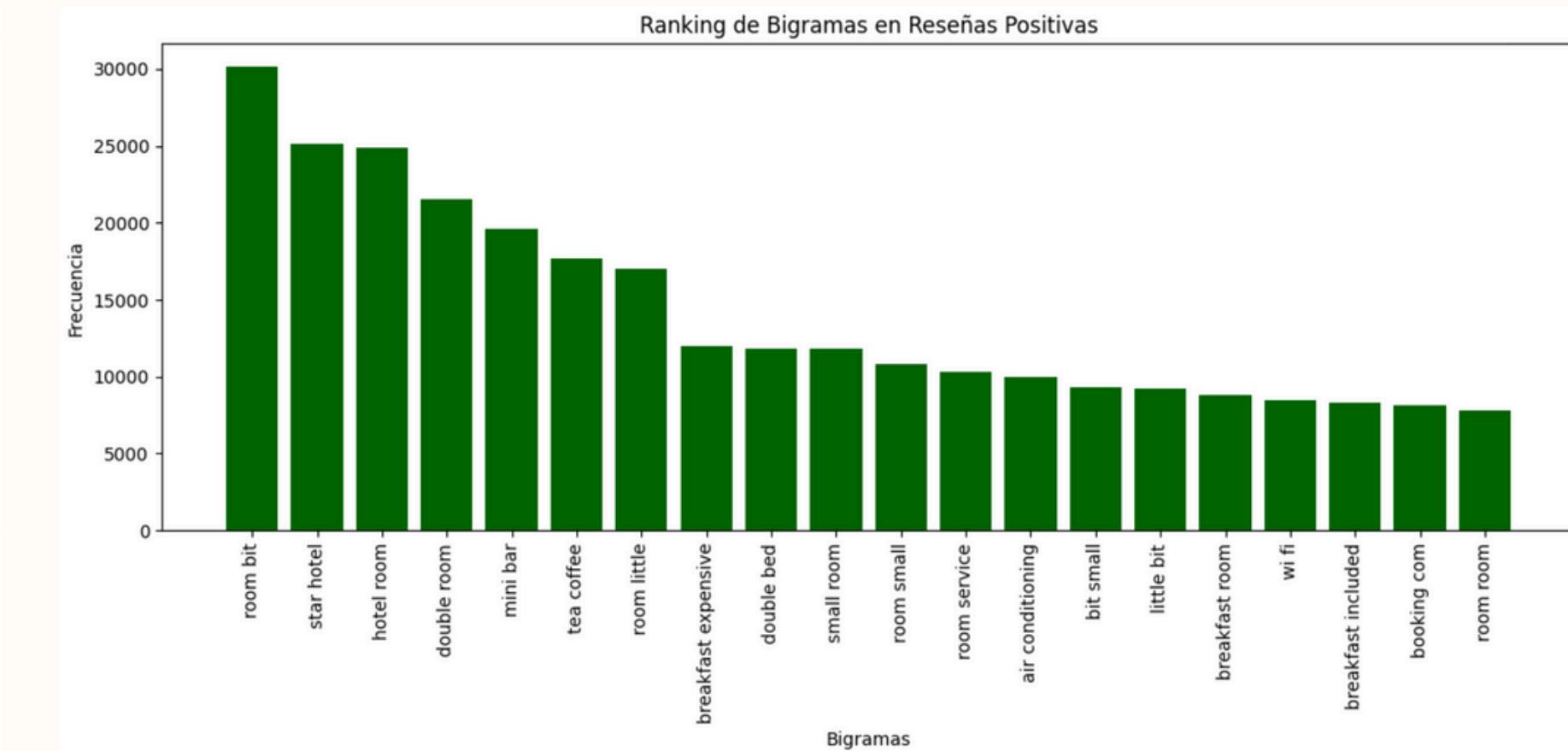
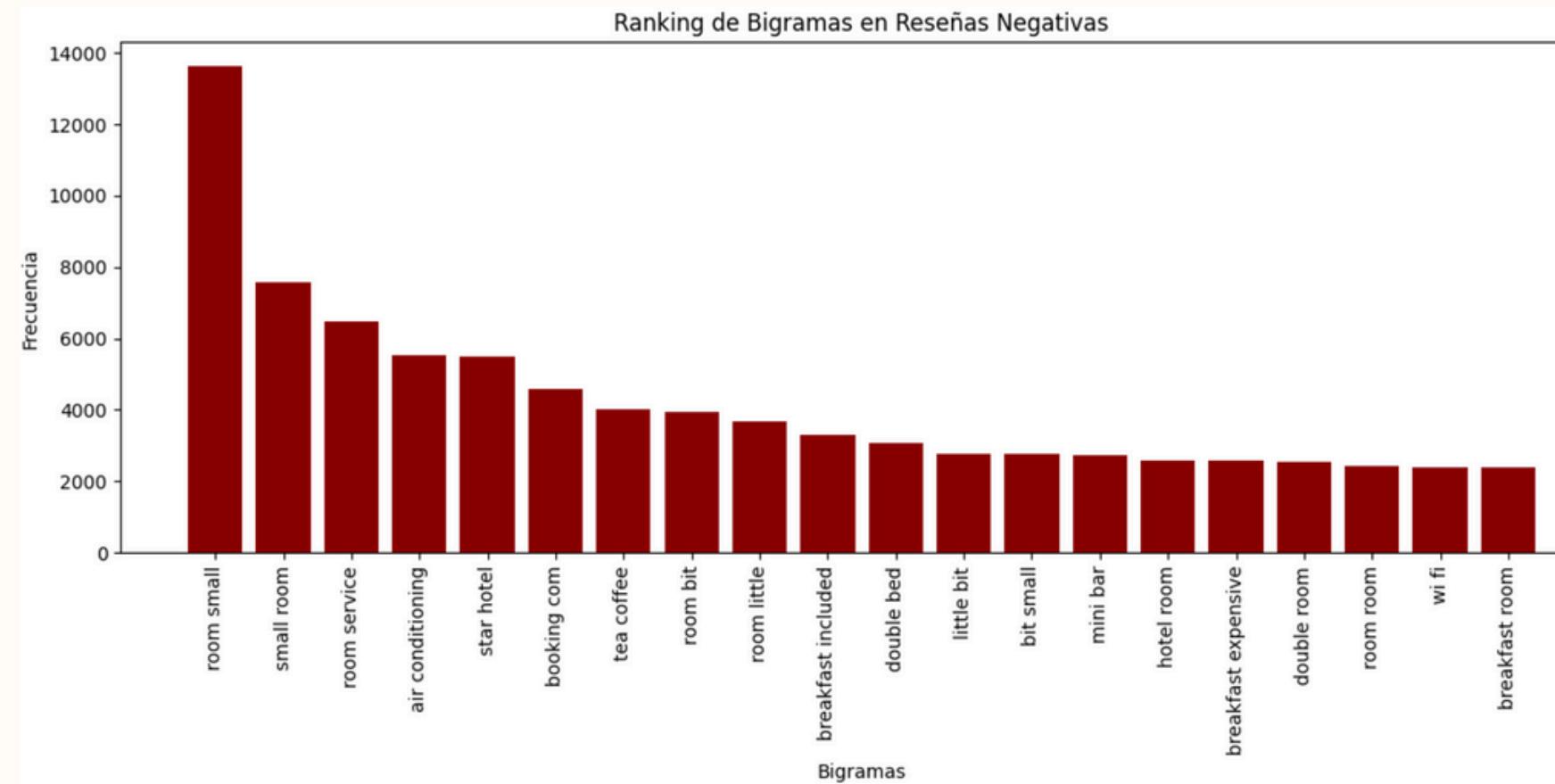
ANÁLISIS CONTEXTUAL

En el análisis anterior la palabra “room” fue una de las más frecuentes, pero por sí misma carece de relevancia sin un contexto. Si realizamos un análisis de contexto sobre las palabras podemos capturar cuáles fueron los “puntos flojos o altos” identificados por los clientes, como por ejemplo el servicio a la habitación o un tamaño de habitación pequeño.



PREPARACIÓN Y ENTRENAMIENTO

CREACION DE CARACTERISTICAS CON PALABRAS CLAVE



Generamos las nuevas características agrupando significados similares

```
#breakfast
breakfast = np.zeros(len(df))
for i in range(len(df)):
    if ("breakfast expensive" in negative_comment[i]) or ("breakfast included" in negative_comment[i])
        or ("breakfast room" in negative_comment[i]) or ("tea coffee" in negative_comment[i]):
        breakfast[i] = -1
np.sum(breakfast)
```

PREPARACIÓN Y ENTRENAMIENTO

CREACION DE CARACTERISTICAS CON TAGS

De los tags también podemos abstraer información importante, estos proporcionan un resumen directo y categorizado de los temas clave mencionados en las reseñas ayudando a identificar rápidamente aspectos recurrentes o prioritarios, facilitan el análisis temático y permiten correlacionar características específicas con las evaluaciones generales.

Algunos de los que identificamos que podrían estar estrechamente relacionados a la percepción y puntuación:

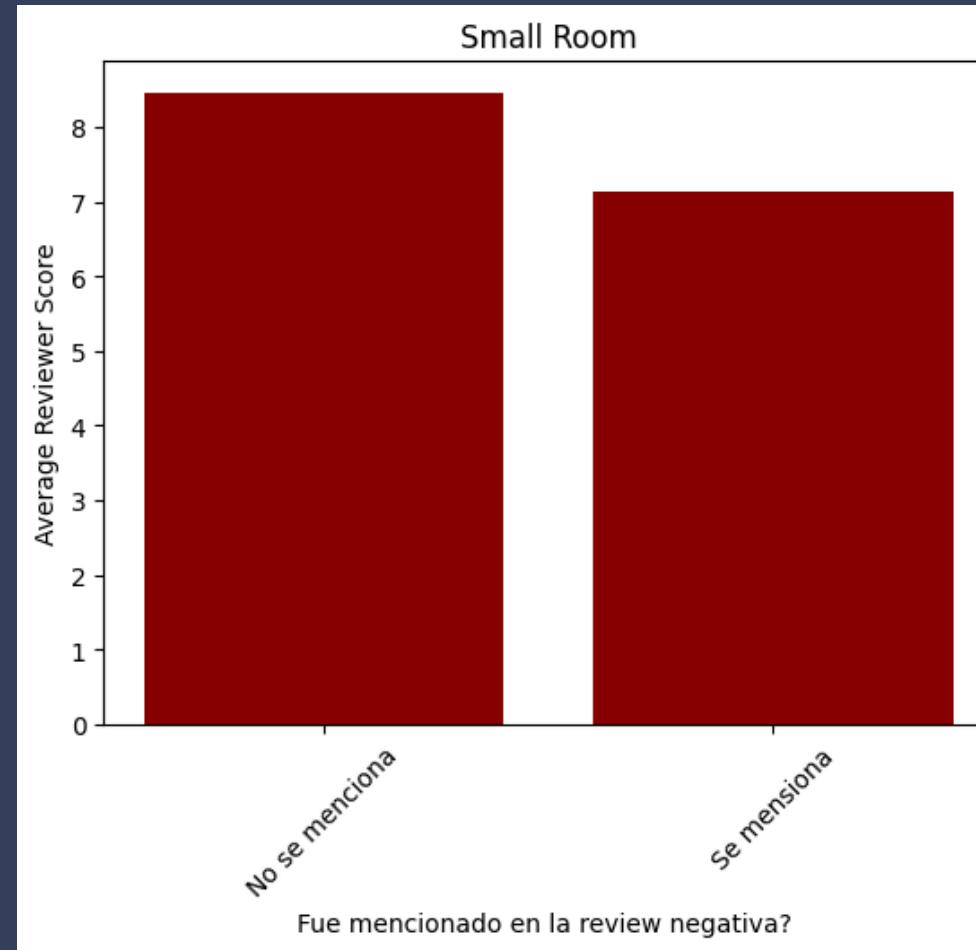
- Tipo de viaje de placer, negocios o ambos
- Viaje en pareja, solo, grupo, familia con niños mayores, Familia con niños pequeños
- Largo de estadía en días

```
# Tipo de viaje de placer, negocios o ambos.  
# "Trip_type": 1 = Placer, 2 = Negocios, 3 = Ambos, 0 = ninguno.  
df['Leisure'] = df['Tags'].map(lambda x: 1 if 'Leisure trip' in x else 0)  
df['Business'] = df['Tags'].map(lambda x: 2 if 'Business trip' in x else 0)  
df['Trip_type'] = df['Leisure'] + df['Business']
```

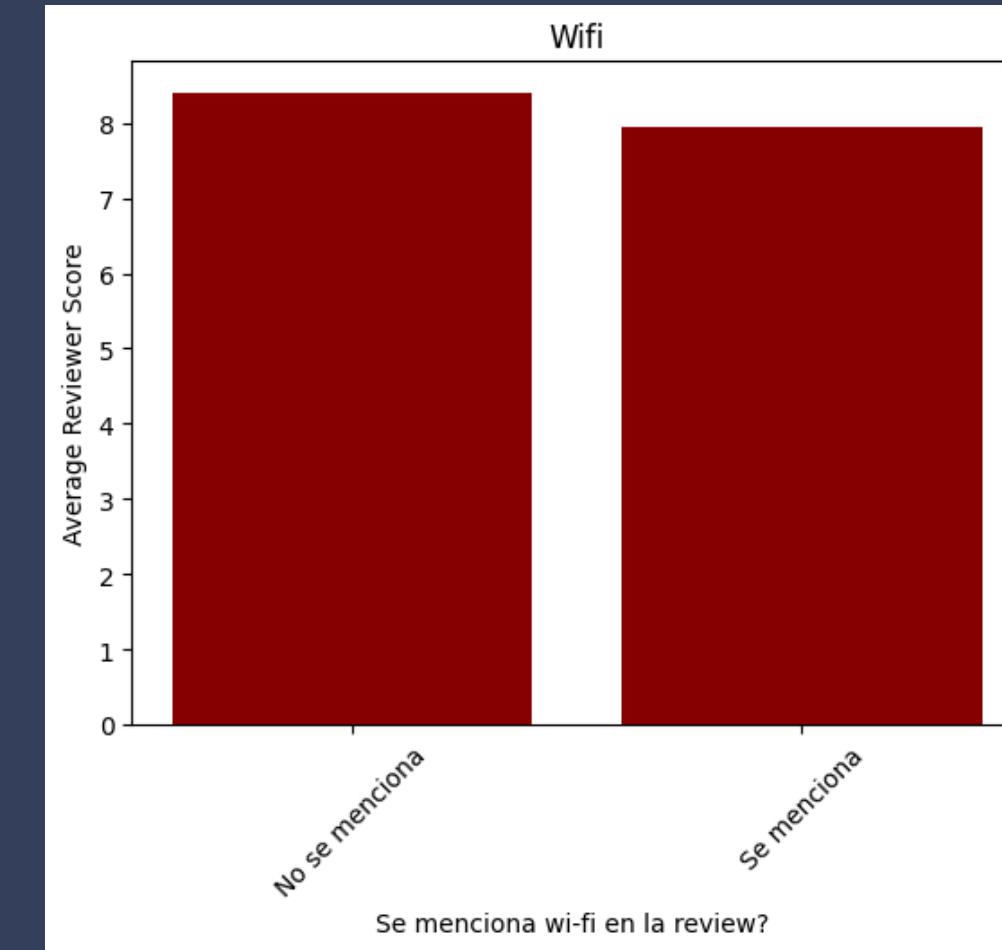
PREPARACIÓN Y ENTRENAMIENTO

SELECCIÓN DE CARACTERÍSTICAS

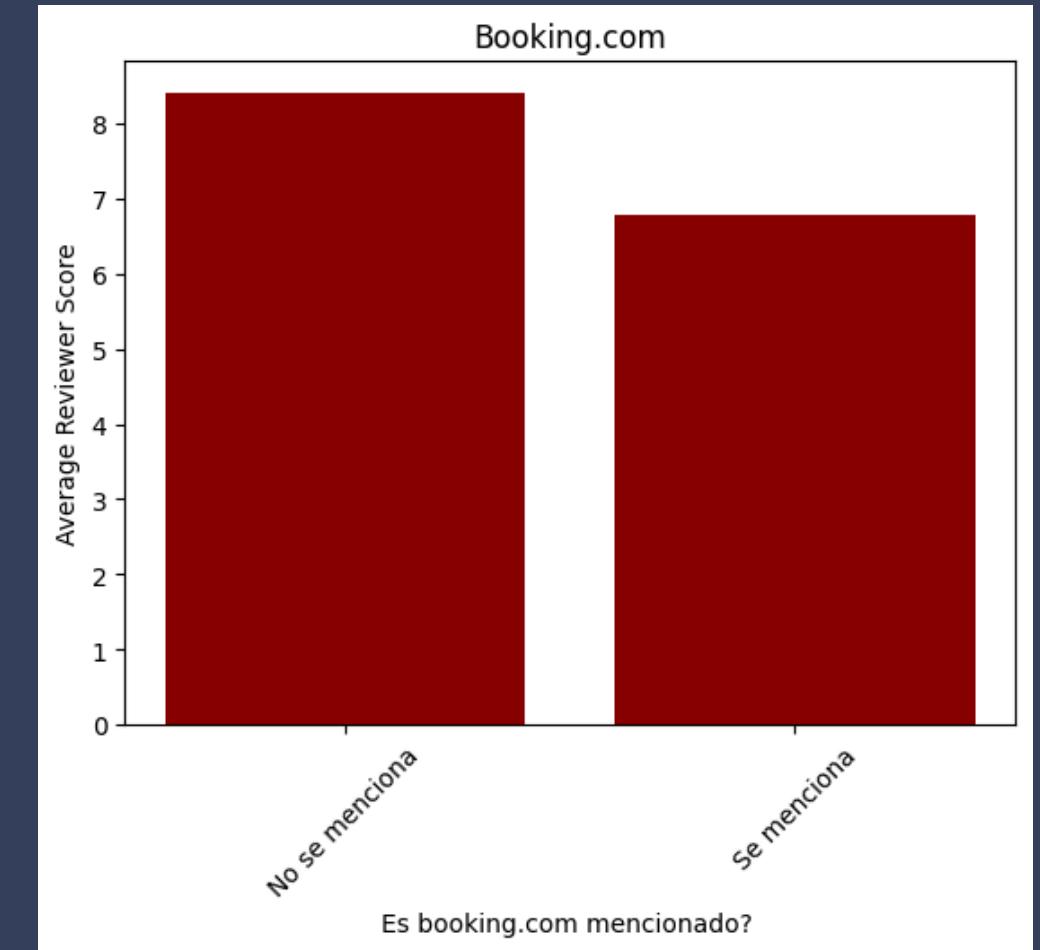
Para revisiones negativas...



Es bastante importante en la calificación del revisor



Cuando se menciona puede significar una puntuación mas baja



Cuando no se menciona booking.com la puntuación es mayor

Otras relevantes: Air Conditioning, Room Problem.

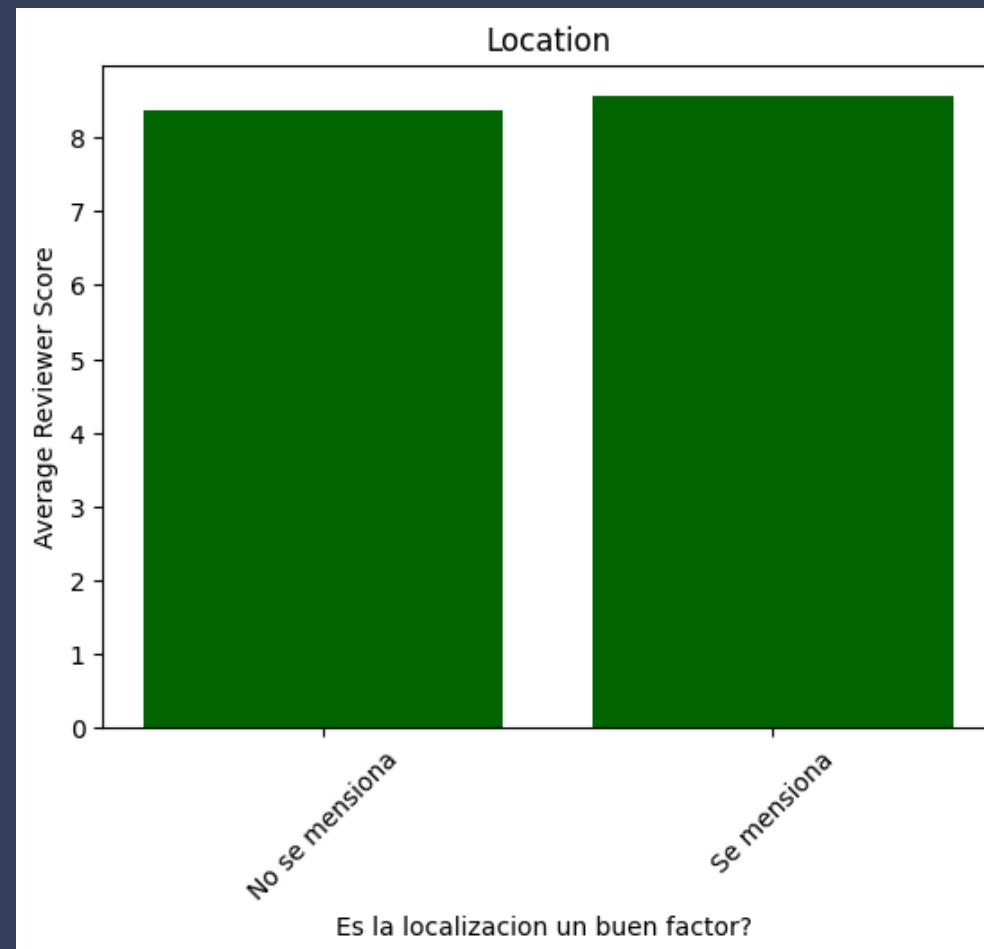
Graficas en notebook adjunto

En cuanto a otras como breakfast no fueron influyentes en ningún caso (negativo o positivo)

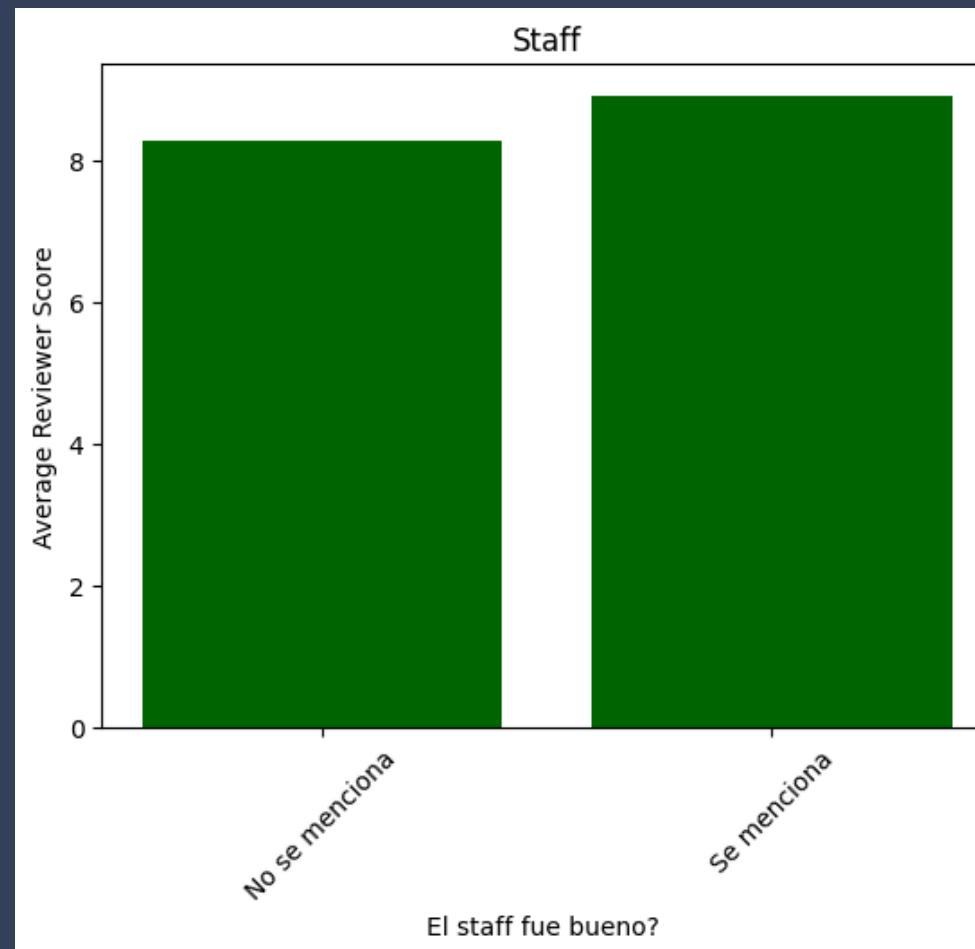
PREPARACIÓN Y ENTRENAMIENTO

SELECCIÓN DE CARACTERÍSTICAS

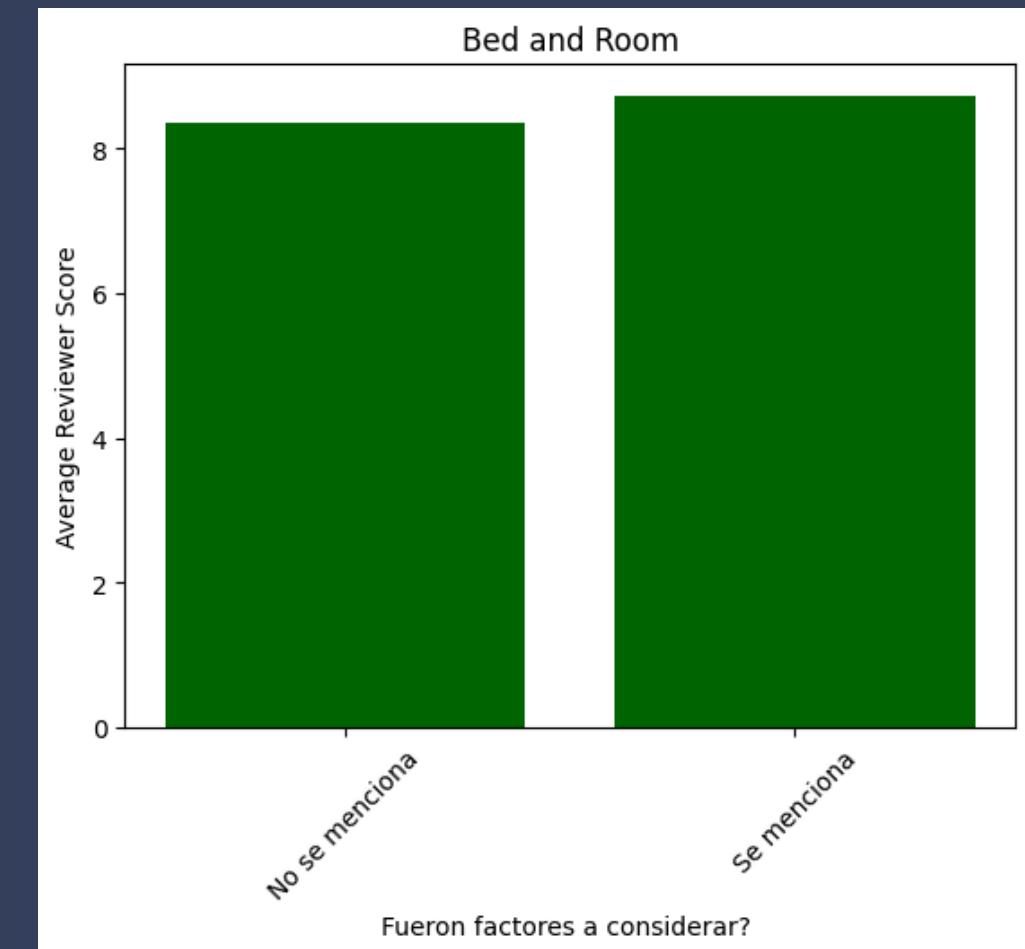
Para revisiones positivas...



No es influyente en la calificación del revisor



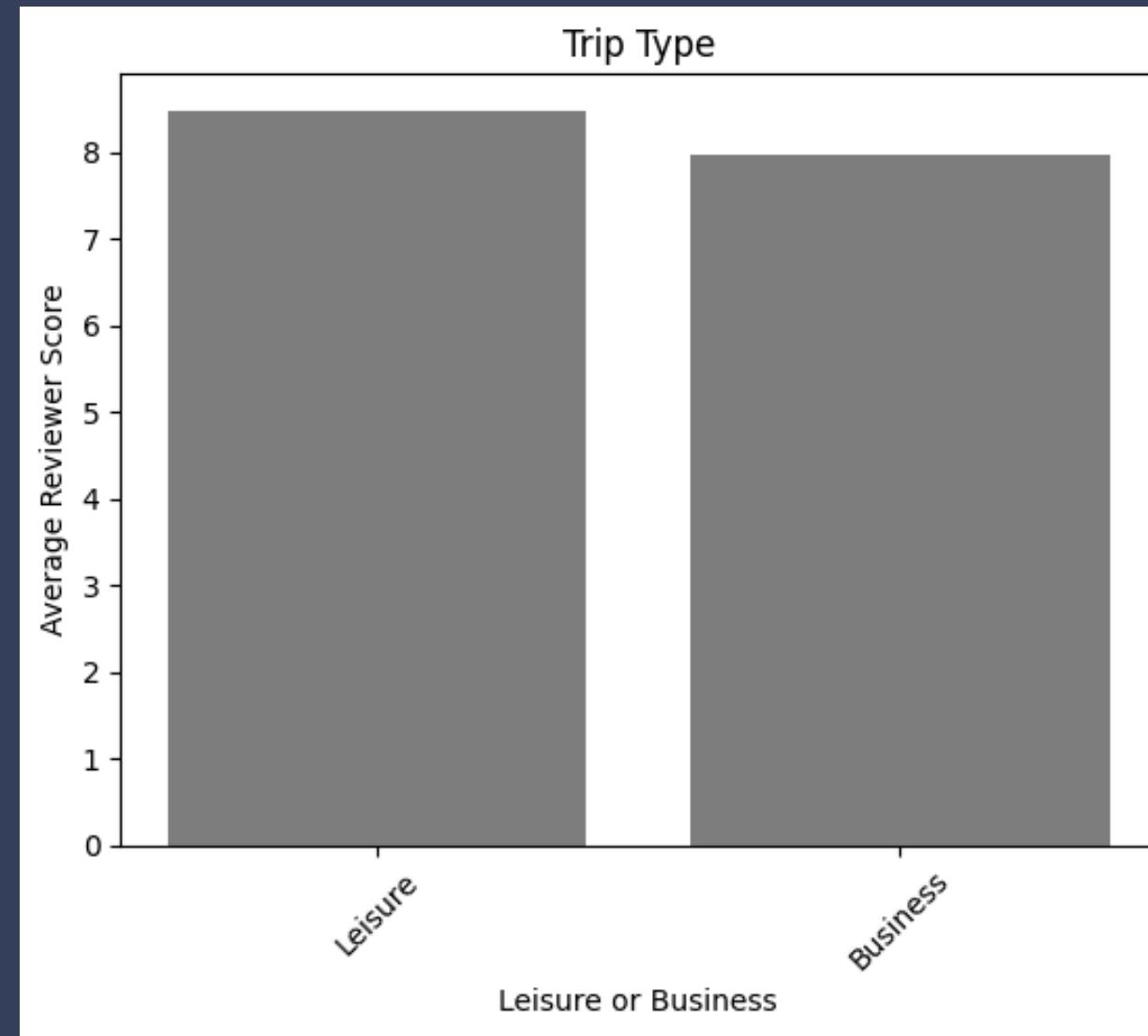
Puede influir y mejorar la calificación



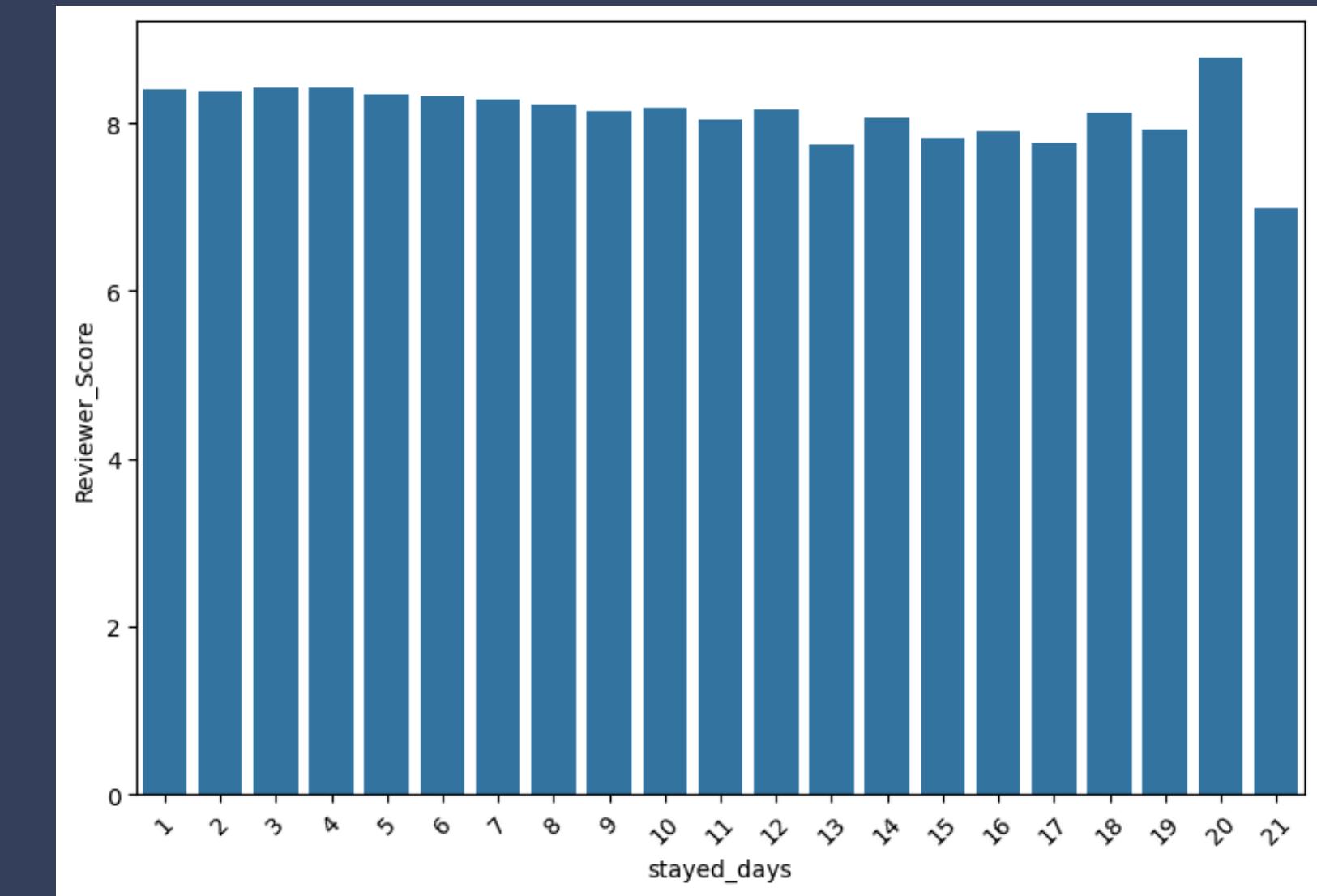
No nos parece importante en la calificación

PREPARACIÓN Y ENTRENAMIENTO

SELECCIÓN DE CARACTERÍSTICAS (TAGS)



Los viajes de placer mejoran gradualmente la calificación



Cuantos mayor es la estadía puede afectar negativamente al score

En el caso de otros tags y campos que analizamos no tuvieron casi injerencia en cuanto al score algunos de ellos fueron:
Estadía individual o compartida, mes de estadía, año de estadía.

PREPARACIÓN Y ENTRENAMIENTO

SELECCIÓN DE MODELO 1

Utilizamos XGBOOST

Nuestra elección como primer modelo fue **XGBoost** debido a que, configurado con regresión gamma, tiene la capacidad para manejar eficientemente datos positivos y asimétricos, características presentes en la distribución de nuestro Dataset. Este modelo combina alta precisión, rapidez de entrenamiento y regularización para evitar el sobreajuste, mientras que la función de regresión gamma optimiza el ajuste en distribuciones con colas largas, permitiendo una predicción más precisa en escenarios donde los datos no siguen una distribución normal.

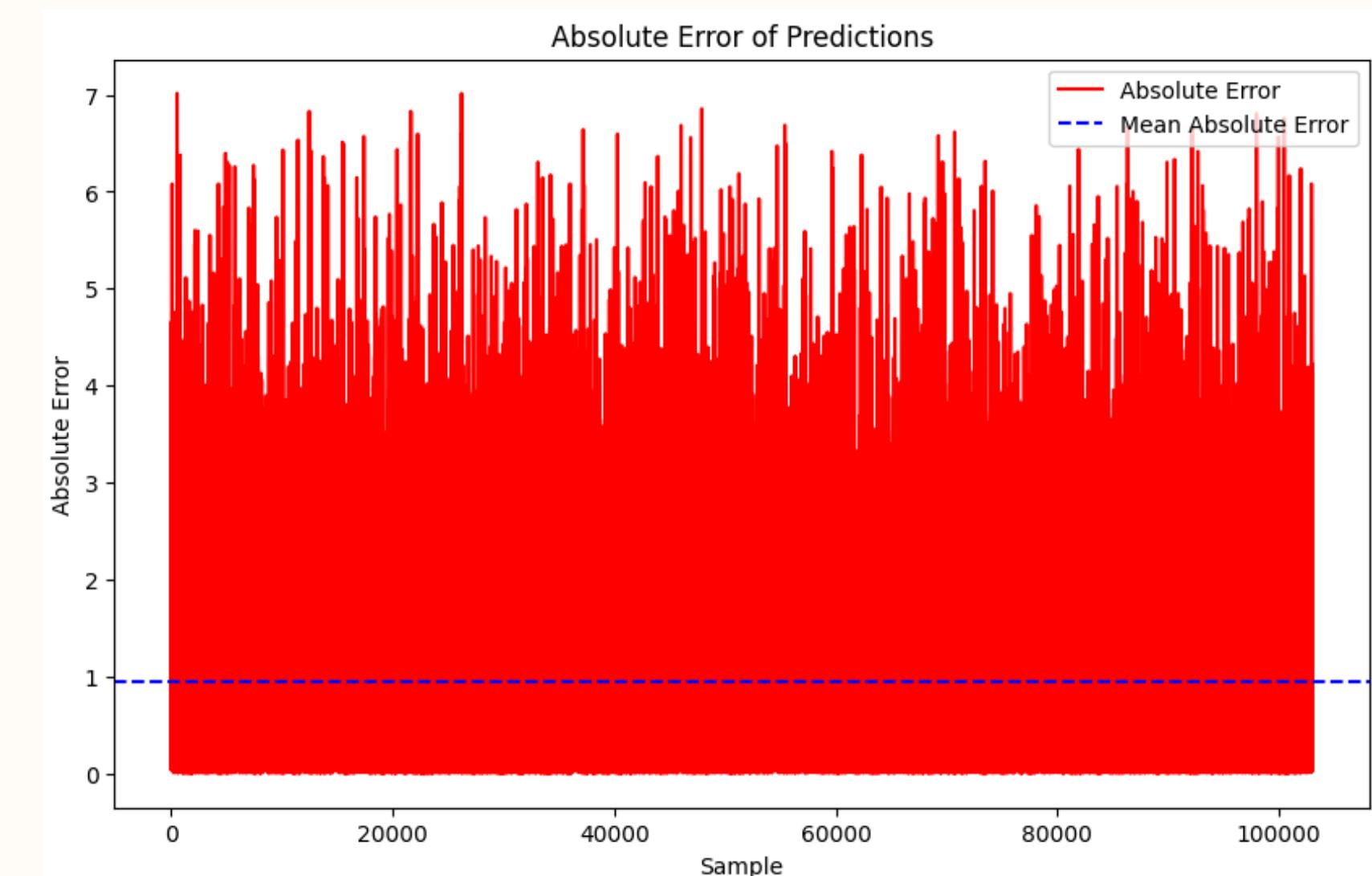
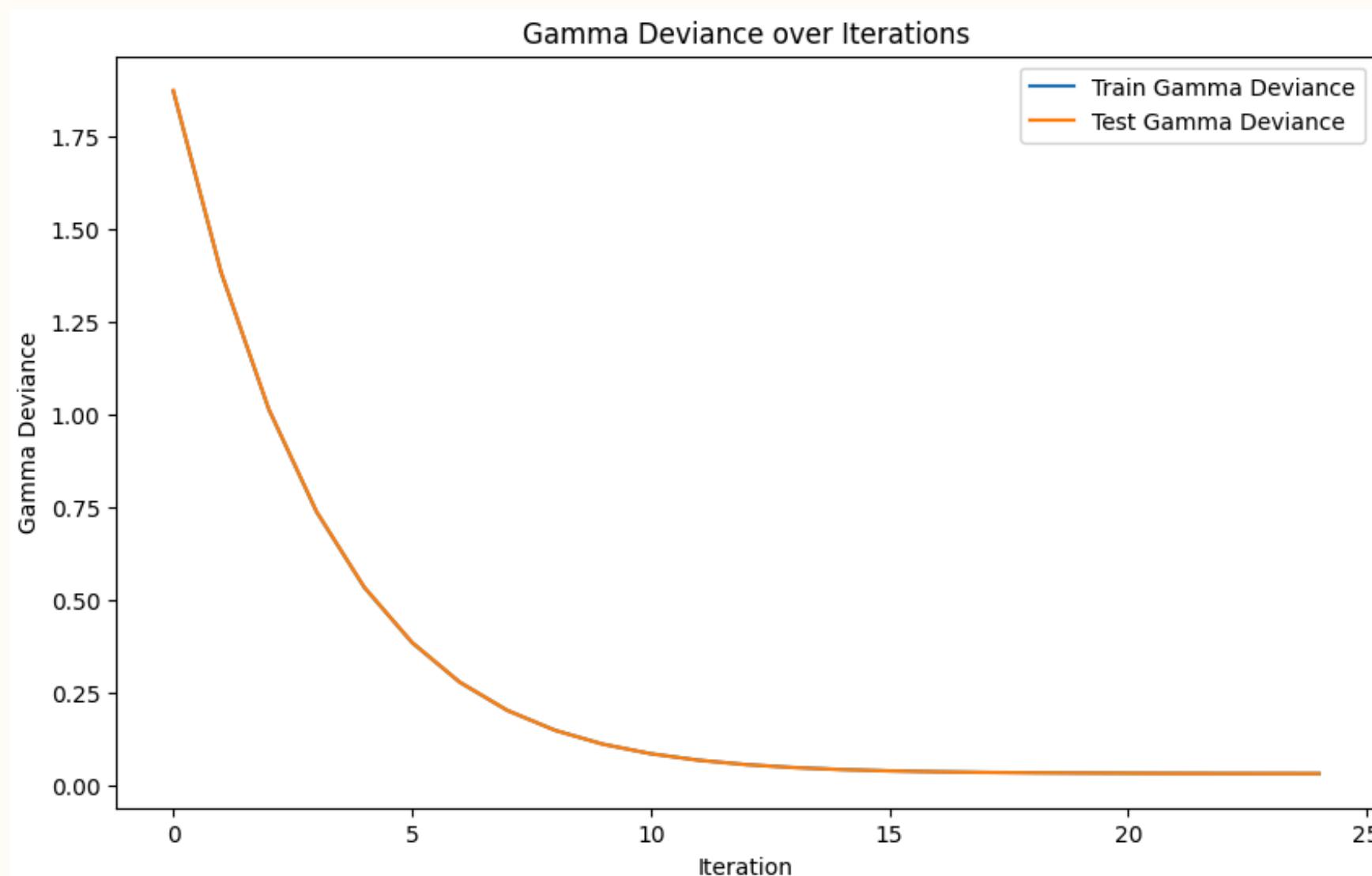


```
xg_test = xgb.DMatrix(X_test, label=y_test)
param = {}
param['objective'] = 'reg:gamma'
param['eta'] = 0.2
param['max_depth'] = 4
param['silent'] = 1
param['nthread'] = 4
watchlist = [(xg_train, 'train'), (xg_test, 'test')]
num_round = 25
# Guardar el historial de errores
history = {} # Para almacenar los resultados de cada iteración

bst = xgb.train(param, xg_train, num_round, watchlist, evals_result=history)
pred = bst.predict(xg_test)
```

MÉTRICAS

En términos generales, el modelo parece estar haciendo un trabajo decente, dado el error promedio relativamente bajo (11.26%). Aunque, la variabilidad del error en las muestras individuales (con valores de error de hasta 7) nos indica que el modelo tiene dificultades para predecir con precisión en algunos casos específicos, podría mejorarse combinando con el segundo método, pero la RAM disponible no lo soportó.



SEGUNDO MÉTODO

Unicamente utiliza el texto de las reviews. Este método está basado en Natural Language Processing (NLP) que consiste en intentar extraer las emociones desde el texto crudo, por lo tanto vamos a trabajar con un solo campo de reviews combinando las positivas y negativas

**Random Forest con NLP y
análisis de sentimiento**



CREACION DE CARACTERISTICAS CON PALABRAS CLAVE

```
#Generamos un df nuevo para trabajar sólo en este modelo  
reviews_df=df  
#Agrupamos las dos reviews en un solo campo  
reviews_df["review"] = reviews_df["Negative_Review"] + reviews_df["Positive_Review"]  
#Agregamos una nueva caracteristica que será el objetivo para este modelo  
reviews_df["is_bad_review"] = reviews_df["Reviewer_Score"].apply(lambda x: 1 if x < 8.0 else 0)  
#Dejamos solo las columnas relevantes  
reviews_df = reviews_df[["review", "is_bad_review"]]  
print(reviews_df.count())
```

Unimos las dos reviews

```
reviews_df["review"] = reviews_df["review"].apply(lambda x: x.replace("No Negative", "").replace("No Positive", ""))
```

Eliminamos los valores por defecto por los falsos positivos

```
# obtenemos una muestra balanceada de reviews_df entre is_bad_review=1 y is_bad_review=0

from sklearn.utils import resample

# Separate majority and minority classes
df_majority = reviews_df[reviews_df.is_bad_review==0]
df_minority = reviews_df[reviews_df.is_bad_review==1]

# Downsample majority class
df_majority_downsampled = resample(df_majority,
    replace=False,      # sample without replacement
    n_samples=len(df_minority),    # to match minority class
    random_state=42) # reproducible results

# Combine minority class with downsampled majority class
reviews_df_old= reviews_df
reviews_df = pd.concat([df_majority_downsampled, df_minority])

# Display new class counts
print(reviews_df.is_bad_review.value_counts())
```

```
reviews_df_presampled=reviews_df
reviews_df_presampled = reviews_df.sample(frac = 0.05, replace = False, random_state=42);
reviews_df_presampled.count()
```

RESAMPLE

Resize porque
colapsa el collab
por falta de RAM

LIMPIEZA DE LA REVIEW FINAL

```
from nltk.corpus import wordnet

def get_wordnet_pos(pos_tag):
    if pos_tag.startswith('J'):
        return wordnet.ADJ
    elif pos_tag.startswith('V'):
        return wordnet.VERB
    elif pos_tag.startswith('N'):
        return wordnet.NOUN
    elif pos_tag.startswith('R'):
        return wordnet.ADV
    else:
        return wordnet.NOUN

import string
import nltk
nltk.download('stopwords')
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize
from collections import Counter
nltk.download('punkt')
nltk.download('averaged_perceptron_tagger')
nltk.download('wordnet')
from nltk import pos_tag
nltk.download('stopwords')
from nltk.corpus import stopwords
from nltk.tokenize import WhitespaceTokenizer
from nltk.stem import WordNetLemmatizer
```

```
def clean_text(text):
    # lower text
    text = text.lower()
    # tokenize text and remove punctuation
    text = [word.strip(string.punctuation) for word in text.split(" ")]
    # remove words that contain numbers
    text = [word for word in text if not any(c.isdigit() for c in word)]
    # remove stop words
    stop = stopwords.words('english')
    text = [x for x in text if x not in stop]
    # remove empty tokens
    text = [t for t in text if len(t) > 0]
    # pos tag text
    pos_tags = pos_tag(text)
    # lemmatize text
    text = [WordNetLemmatizer().lemmatize(t[0], get_wordnet_pos(t[1])) for t in pos_tags]
    # remove words with only one letter
    text = [t for t in text if len(t) > 1]
    # join all
    text = " ".join(text)
    return(text)

# clean text data
reviews_df["review_clean"] = reviews_df["review"].apply(lambda x: clean_text(x))
```

Agregamos analisis de sentimiento usando Vader

```
from nltk.sentiment.vader import SentimentIntensityAnalyzer
nltk.download('vader_lexicon')
#Usamos vader (NLTK) para poder analisar sentimiento y agregarlo al df
sid = SentimentIntensityAnalyzer()
reviews_df["sentiments"] = reviews_df["review"].apply(lambda x: sid.polarity_scores(x))
reviews_df = pd.concat([reviews_df.drop(['sentiments'], axis=1), reviews_df['sentiments'].apply(pd.Series)], axis=1)
```

Vectorizamos frases mediante genesim

```
#Usamos genesim para obtener vectores basados en review_clean
from gensim.test.utils import common_texts
from gensim.models.doc2vec import Doc2Vec, TaggedDocument

documents = [TaggedDocument(doc, [i]) for i, doc in enumerate(reviews_df["review_clean"].apply(lambda x: x.split(" ")))]

# entrenamos un modelo Doc2Vec con la info del df corregido
model = Doc2Vec(documents, vector_size=5, window=2, min_count=1, workers=4)

doc2vec_df = reviews_df["review_clean"].apply(lambda x: model.infer_vector(x.split(" "))).apply(pd.Series)
doc2vec_df.columns = ["doc2vec_vector_" + str(x) for x in doc2vec_df.columns]
reviews_df = pd.concat([reviews_df, doc2vec_df], axis=1)
```

```
#Usamos tf-idf para vectorizar las palabras de las reviews con analisis de contexto
from sklearn.feature_extraction.text import TfidfVectorizer
tfidf = TfidfVectorizer(min_df = 10)
tfidf_result = tfidf.fit_transform(reviews_df["review_clean"]).toarray()
tfidf_df = pd.DataFrame(tfidf_result, columns = tfidf.get_feature_names_out())
tfidf_df.columns = ["word_" + str(x) for x in tfidf_df.columns]
tfidf_df.index = reviews_df.index
reviews_df = pd.concat([reviews_df, tfidf_df], axis=1)
```

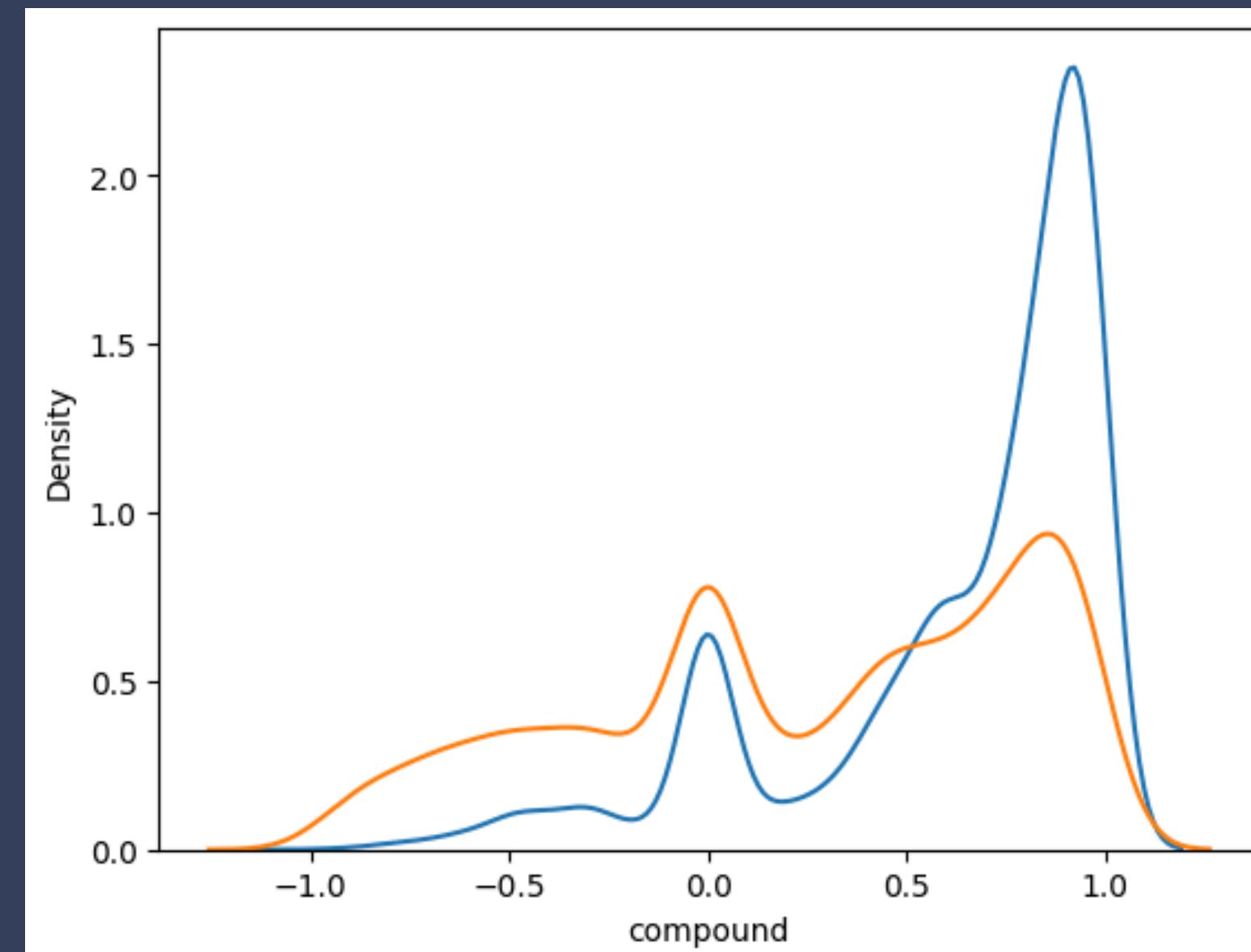
	review	pos
80396	Lovely clean comfortable warm	1.000
228502	Helpful friendly great staff	0.910
384497	Great location comfort clean	0.903
145743	Good value amazing location	0.901
389173	Great location Clean safe comfortable friendl...	0.890
436901	Lovely comfortable rooms	0.877

Vectorizamos palabras segun contexto con tf-idf

Con esto podemos tener frases con significacion sentimental (con algunos detalles, al quitar la puntuación)

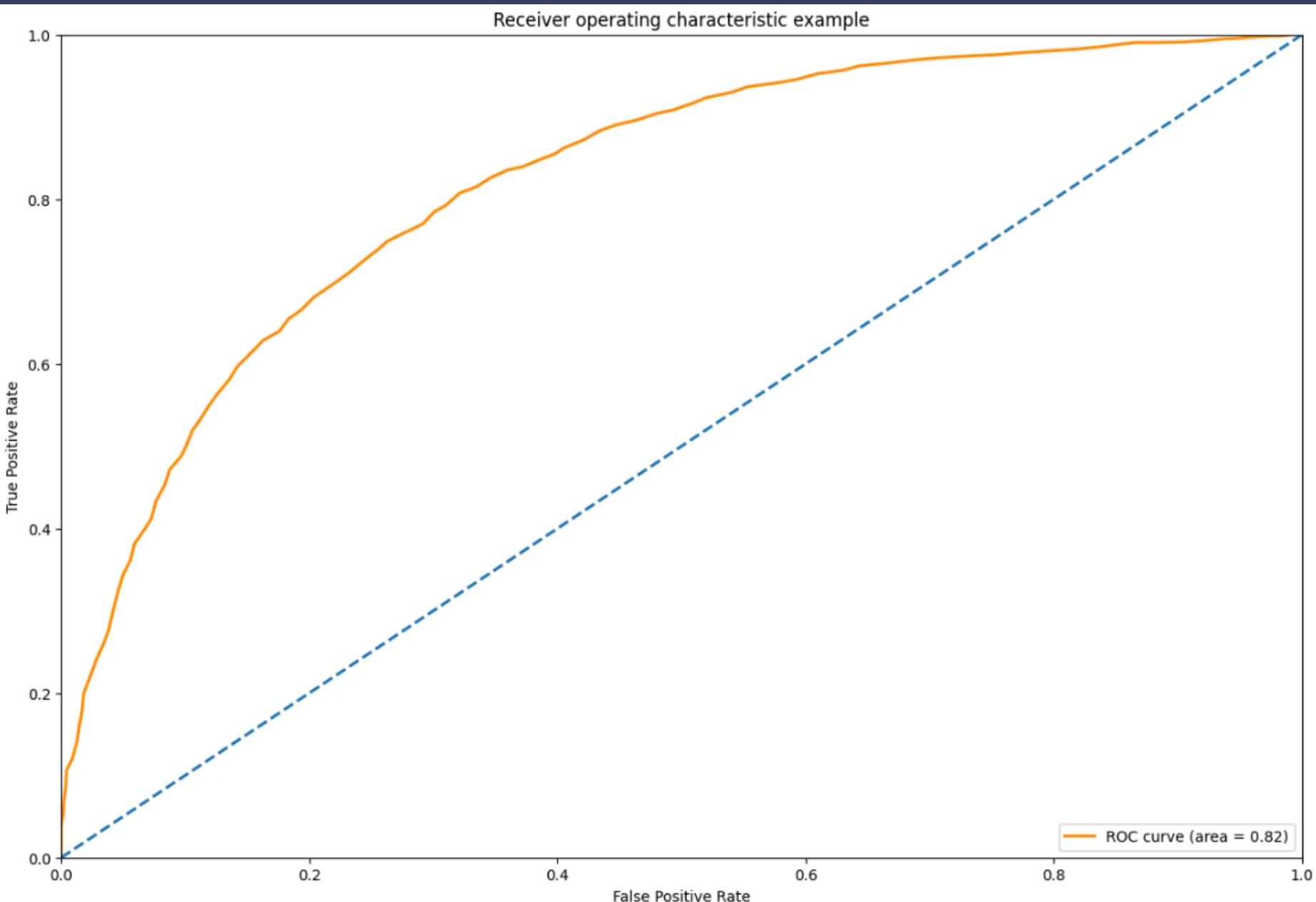
NUEVO ANALISIS SOBRE LAS CARACTERÍSTICAS

El análisis directo sobre las palabras indica que necesitamos sumar el análisis de sentimiento para mejorar la precisión

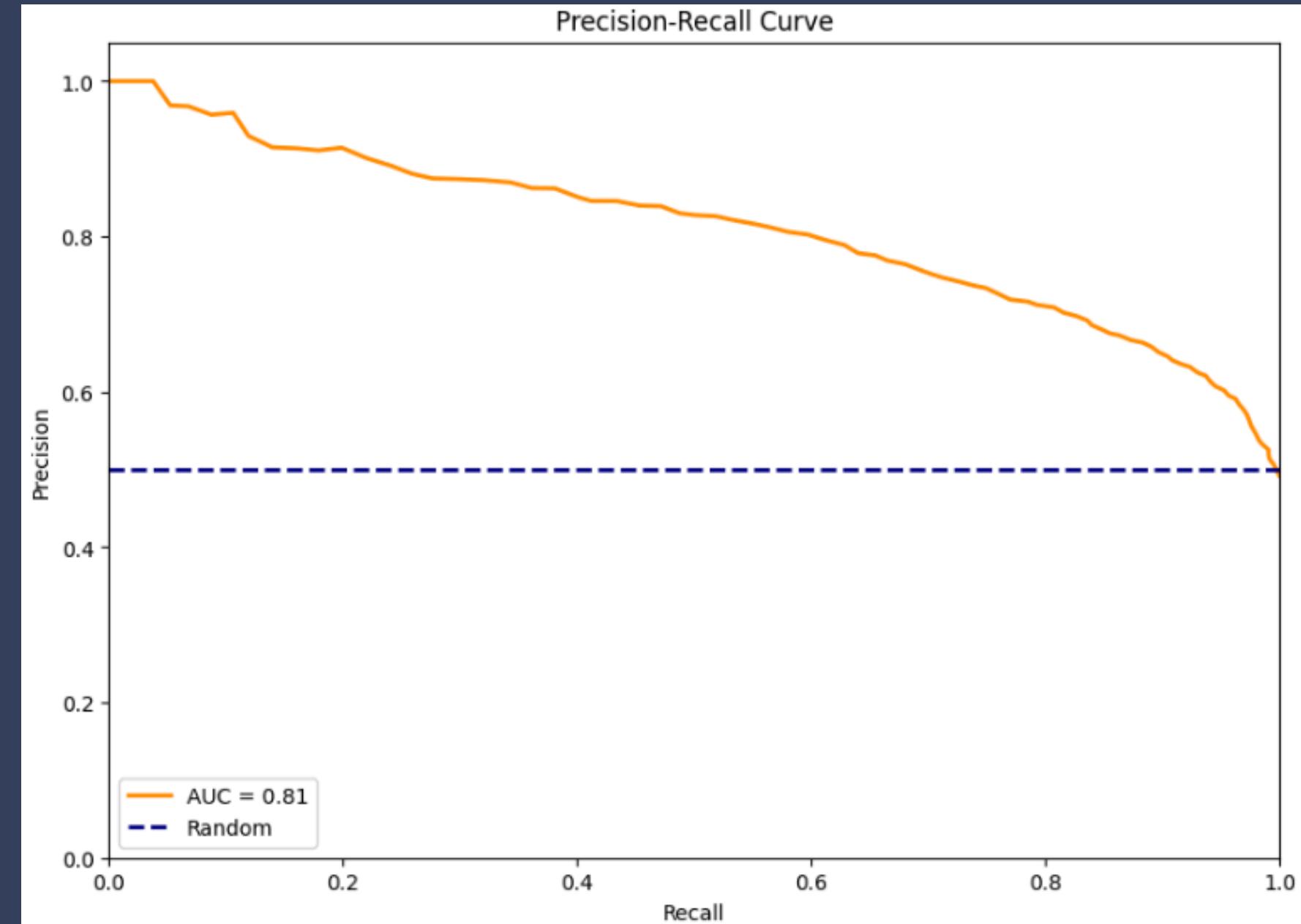


MÉTRICAS

CURVA ROC



CURVA PR



Ambas métricas nos permitieron observar el desempeño del modelo de clasificación. En la curva ROC, el área bajo la curva (AUC = 0.82) indica un buen equilibrio entre la tasa de verdaderos positivos y la tasa de falsos positivos. Por otro lado, la curva PR muestra un AUC de 0.81, lo cual refleja que el modelo tiene una precisión sólida en relación con el recall, aunque se podría mejorar en general.

DEL ANÁLISIS A LA PERSONALIZACIÓN

ANÁLISIS DE SENTIMIENTO

+

CARACTERÍSTICAS DEL HOTEL

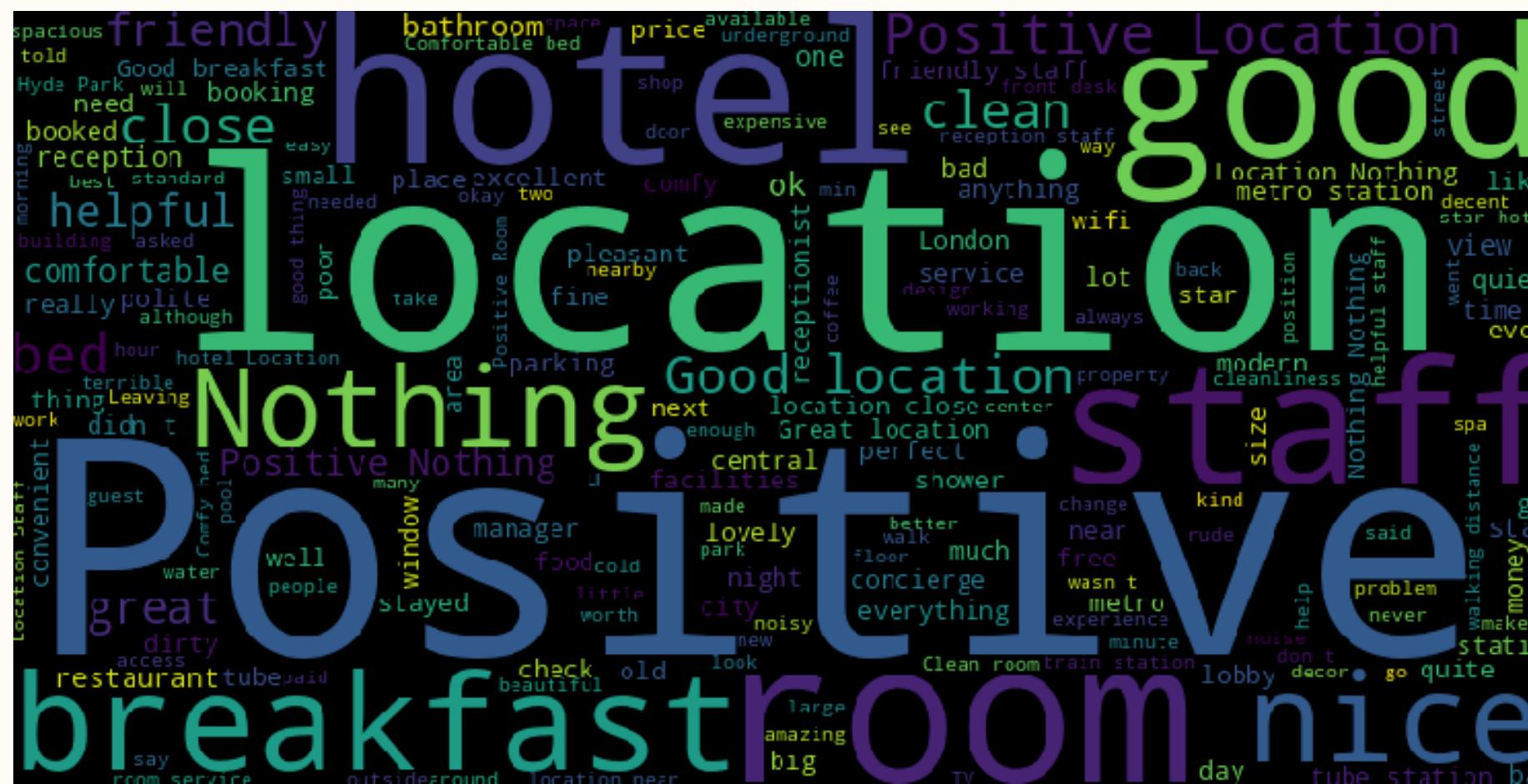
+

HISTORIAL DEL HUESPED Y DEL HOTEL



ANÁLISIS DEL SENTIMIENTO

PRIMERO DESGLOSAMOS LAS RESEÑAS



PARA ENTENDER QUÉ PALABRAS Y FRASES REFLEJAN EXPERIENCIAS POSITIVAS O NEGATIVAS



USO DE LAS ETIQUETAS DE LOS HOTELES

CADA HOTEL TIENE ETIQUETAS QUE DESCRIBEN SUS CARACTERÍSTICAS GENERALES ASÍ COMO LAS DE SUS HABITACIONES ESPECIALES, COMO ‘DELUXE’, ‘BALCÓN’, O ‘CERCANÍA AL CENTRO’. AL VINCULAR ESTO CON LAS RESEÑAS, PODEMOS IDENTIFICAR QUÉ ASPECTOS VALORAN MÁS LOS HUÉSPEDES

HISTORIAL DEL HUESPED Y DEL HOTEL



Los resultados previos del modelo para ese huésped y ese hotel nos ayudan a entender patrones.

¿Qué aspectos tiende a valorar más cada huésped recurrente?

¿Qué genera calificaciones altas en un hotel en particular?"

SISTEMA DE RECOMENDACIONES

¿CÓMO FUNCIONA?

HUESPED CON
PREFERENCIAS
Y QUEJAS

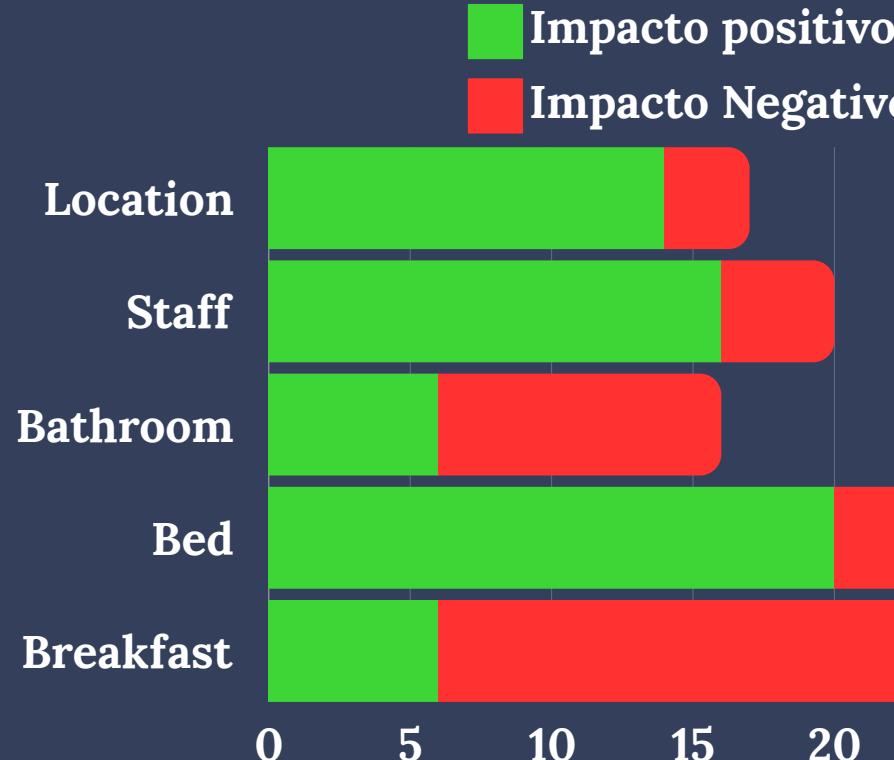
ANALISIS DE
SENTIMIENTO Y
DATOS HISTORICOS

RECOMENDACIÓN
PERSONALIZADA

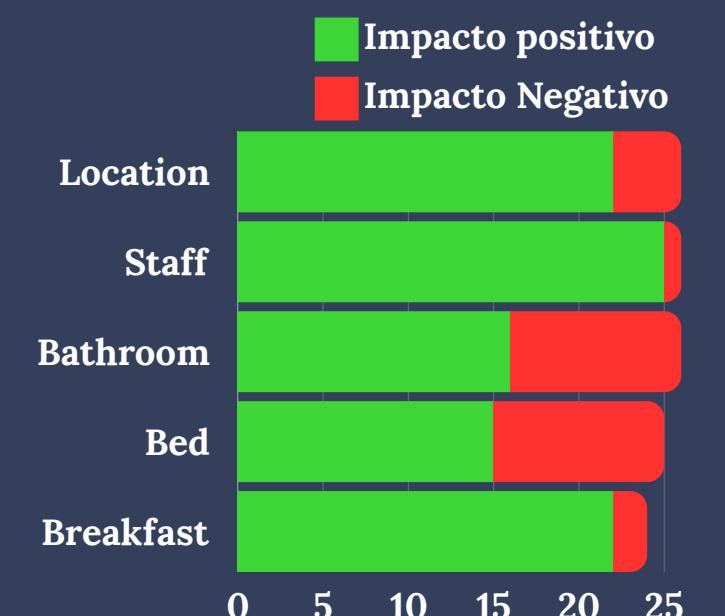
SISTEMA DE RECOMENDACIONES

Cliente: Anne - elije Paris como Destino, y priorizamos en base a los aspectos clave del cliente, comparando con los resultados históricos de los hoteles disponibles en ese lugar

Tiene 6 reseñas cargadas en las cuales sus palabras clave de mayor peso son:

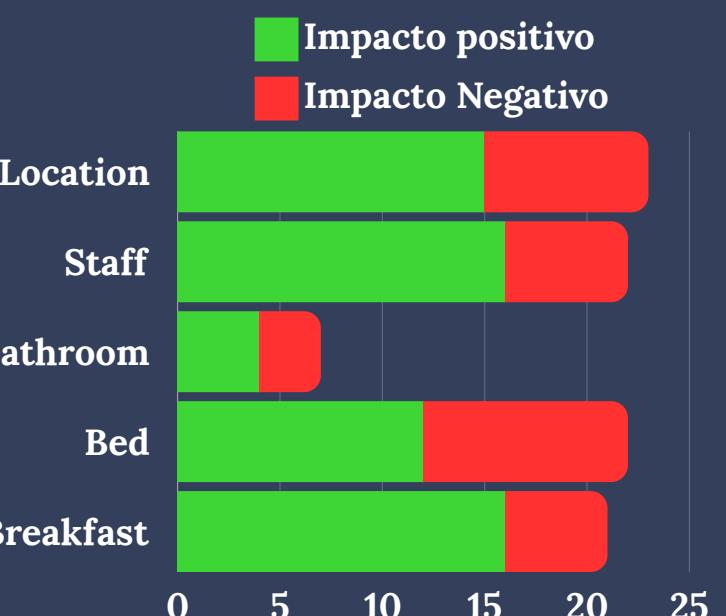


La Tamise Esprit - 9.68

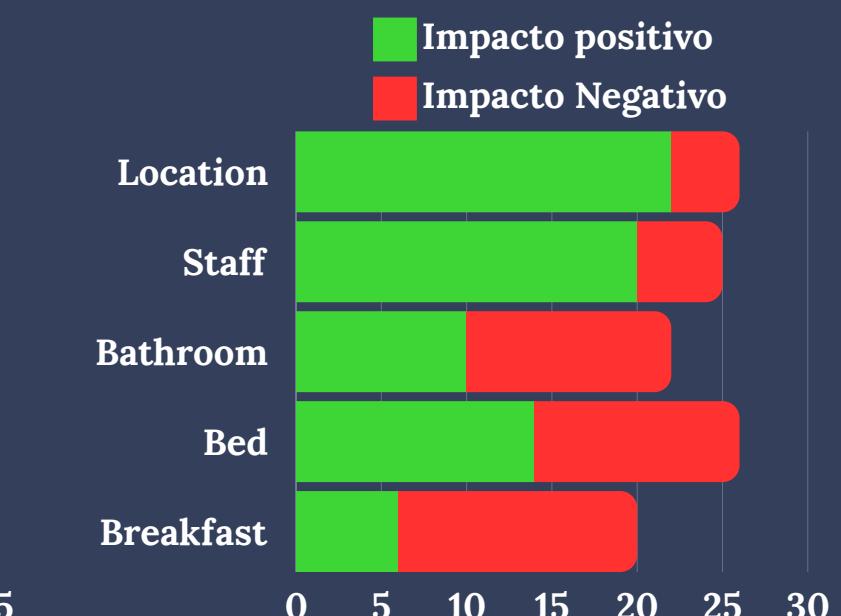


Orden actual (por Score)
 1º Ritz Paris
 2º La Tamise Esprit
 3º Hotel Eiffel Blomet

Ritz Paris - 9.725



Hotel Eiffel Blomet - 9.64



Nuevo orden
 1º La Tamise Esprit
 2º Ritz Paris
 ...
 5º Hotel Eiffel Blomet

NO DEJAR A NADIE AFUERA

APLICAR SIMPLEMENTE LAS RECOMENDACIONES PERSONALIZADAS PREMIA A LOS MEJORES HOTELES PERO CASTIGA A LOS DE PEOR SCORE

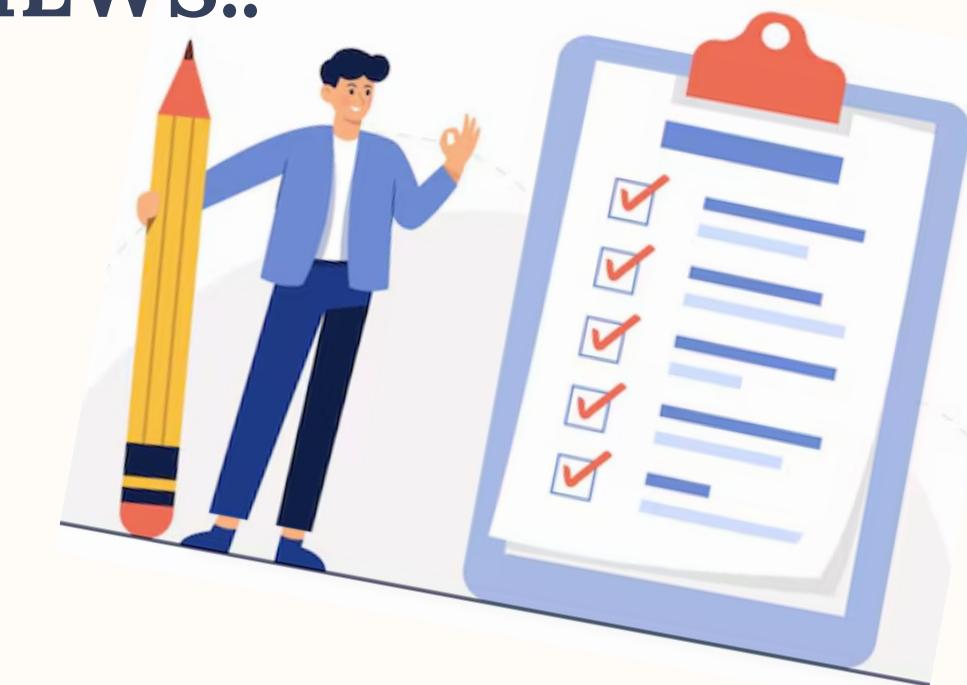
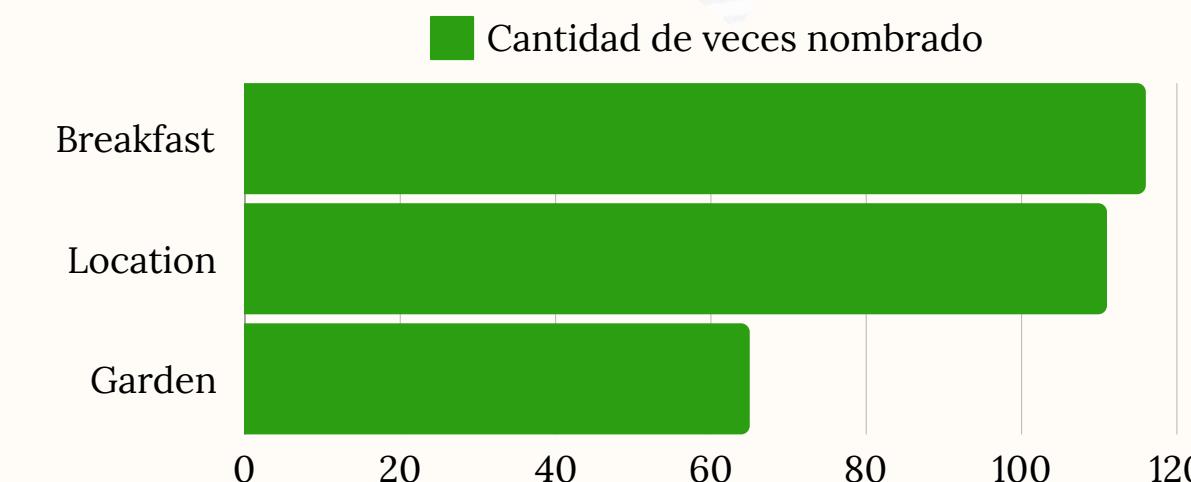
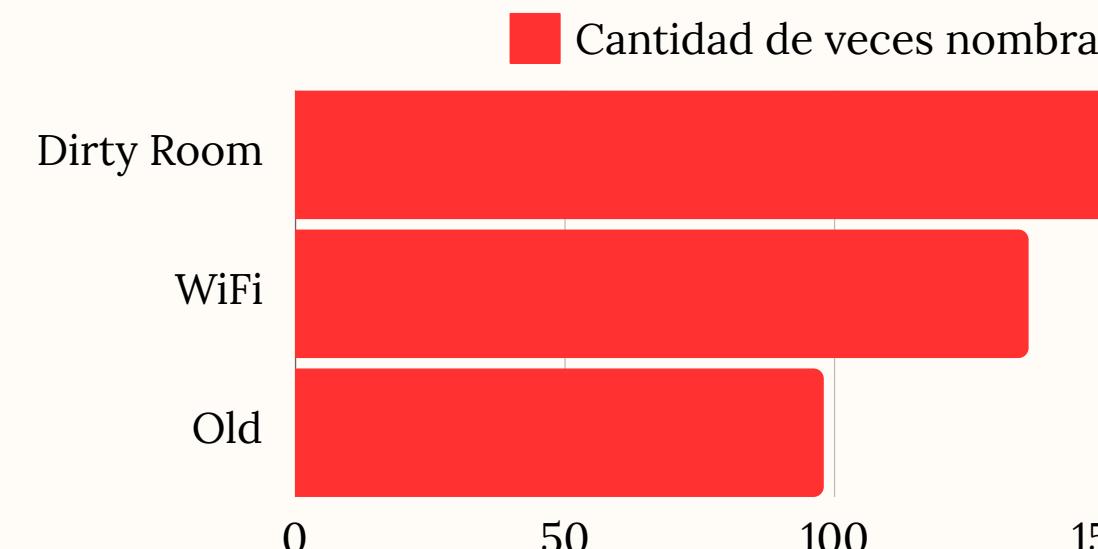
NUESTRA PROPUESTA INCLUYE UN SISTEMA COMPLEMENTARIO QUE PERMITA IDENTIFICAR QUE ASPECTOS SON LOS QUE MÁS INFLUYEN EN LAS NOTAS BAJAS

ESTO PERMITE HACER RECOMENDACIONES SOBRE EN QUÉ ASPECTOS **MEJORAR** Y EN CUALES **ENFATIZAR**

HOTEL LIBERTY

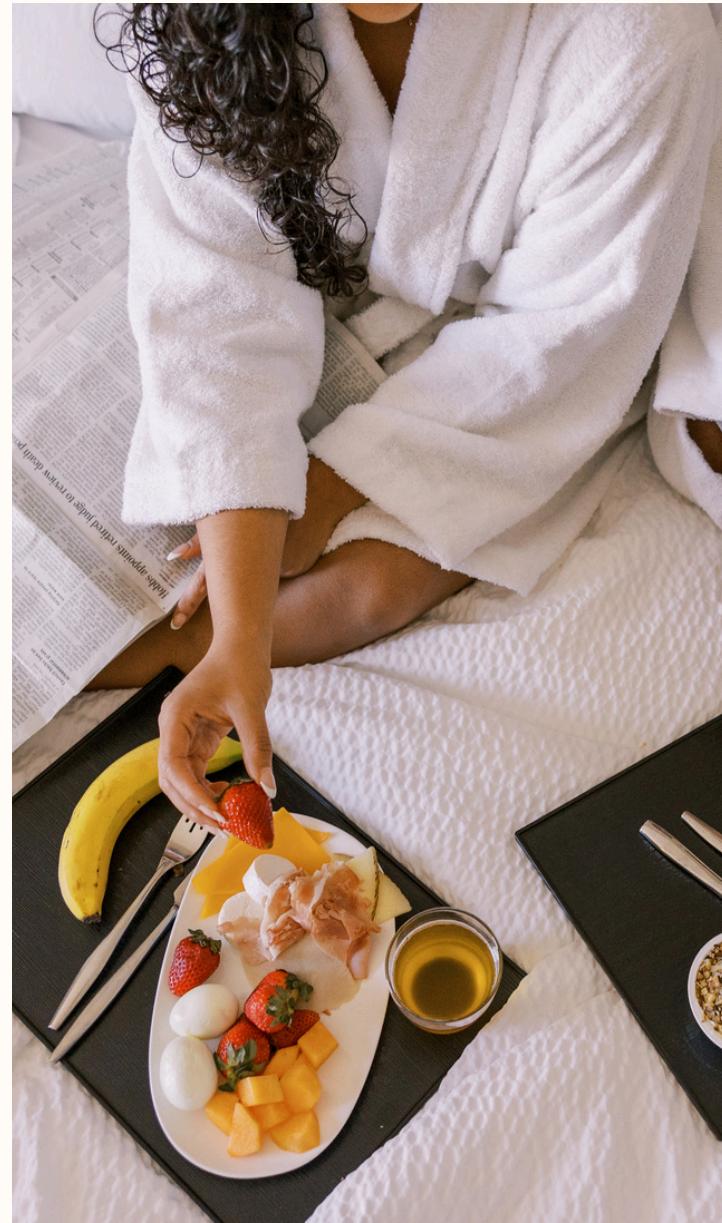
ES EL HOTEL CON EL PEOR PROMEDIO DE RESEÑAS

NUESTRO MODELO IDENTIFICÓ QUE DE 465 REVIEWS..



LLAMADA A LA ACCIÓN

TRANSFORMAR DATOS EN
EXPERIENCIAS MEMORABLES
ES EL FUTURO DE LA
HOSPITALIDAD



RESULTADOS ESPERADOS

01

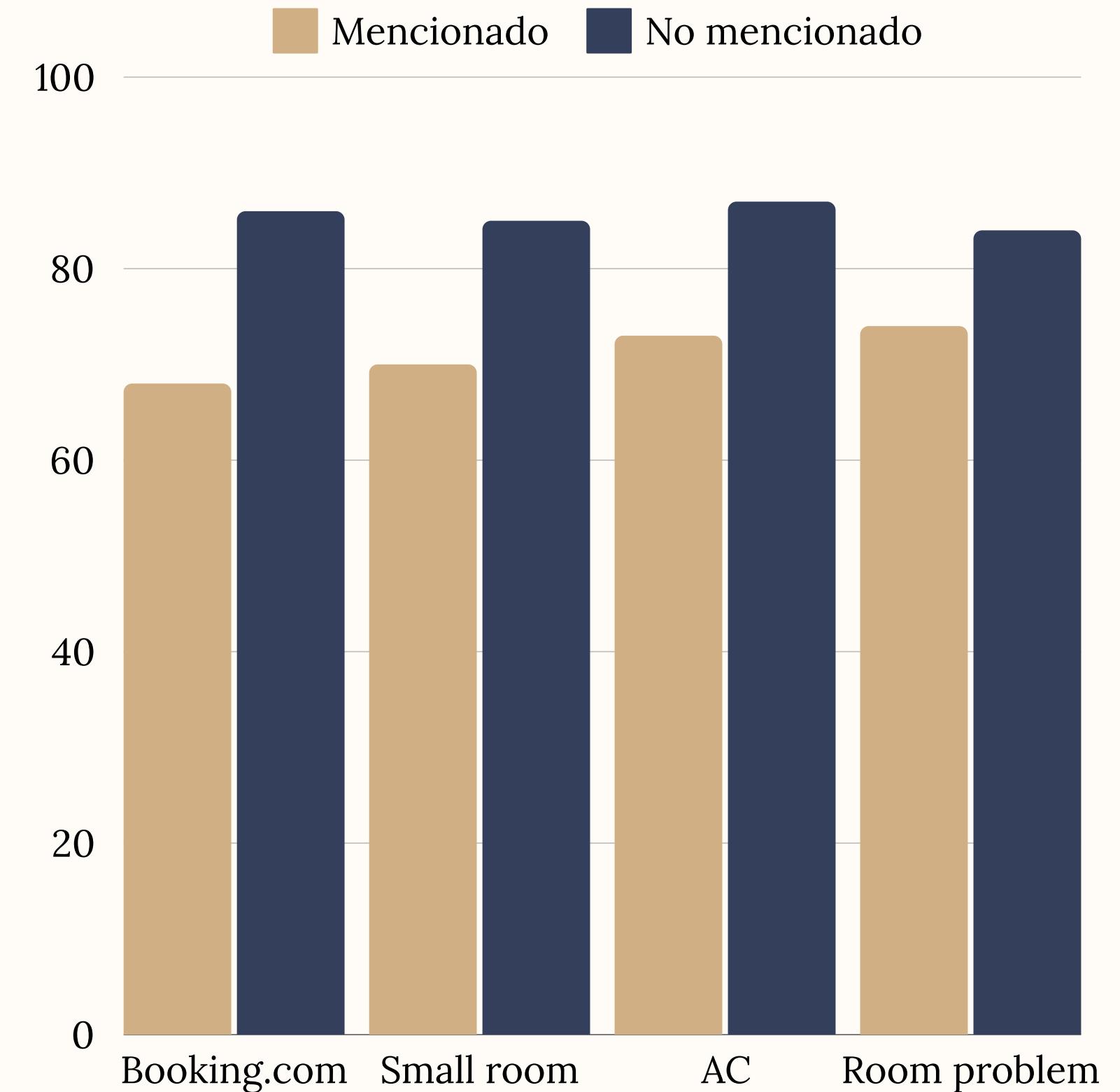
MEJORAR LA EXPERIENCIA DE LOS
HUESPEDES

02

AUMENTAR LA FIDELIDAD

03

OPTIMIZAR LAS OPERACIONES HOTELERAS



MUCHAS
GRACIAS

