DATABRIDGE

SMART SOLUTIONS

CONTEXTO 

La opinión de los usuarios se ha convertido en un dato invaluable en la planificación de estrategias comerciales. Plataformas de reseñas como **Yelp** y **Google Maps** proporcionan una gran cantidad de información sobre la percepción de los usuariosrespecto a diversos negocios, incluyendo restaurantes, hoteles y otros servicios relacionados. Esta retroalimentación es esencial para las empresas, ya que les permite **evaluar su desempeño, identificar áreas de mejora y comprender cómo son percibidas por los usuarios**.
Como parte de una consultora de data, se nos ha contratado para realizar un análisis detallado de la opinión de los usuarios en **Yelp** y **Google Maps** sobre negocios relacionados con el turismo y ocio en el mercado estadounidense.

DESCRIPCION 

Nuestro proyecto consiste en recopilar, depurar y analizar datos de reseñas de Yelp y Google Maps, utilizando técnicas de análisis de sentimientos y machine learning para predecir los rubros de los negocios que experimentarán un crecimiento o declive.
Además, se busca determinar las ubicaciones más adecuadas para establecer nuevos locales comerciales y desarrollar un sistema de recomendación de negocios basado en las preferencias de los usuarios en ambas plataformas.
Aunque nos enfocaremos principalmente en restaurantes, la metodología puede aplicarse a otros tipos de comercios.

OBJETIVO

El objetivo principal del proyecto es brindar a nuestro cliente, parte de un conglomerado de empresas de restaurantes y afines, un **análisis exhaustivo de la opinión de los usuarios en Yelp y Google Maps**. Esto permitirá **identificar tendencias, predecir el crecimiento o decaimiento de rubros comerciales y tomar decisiones estratégicas informadas para mejorar decisiones de gestion e inversion de negocios**. Además se buscará tener un **sistema de recomendación de restaurantes** para los usuarios de ambas plataformas para darle, al usuario por ejemplo la posibilidad de poder conocer nuevos negocios basado en sus gustos y experiencias previas.

ALCANCE

Recopilación de datos: Extracción de datos de reseñas de Yelp y Google Maps, considerando información como *ubicación de los comercios, categorías, puntaje promedio, estado de apertura, usuarios, reseñas realizadas, votos recibidos*, entre otros.

Depuración y almacenamiento de datos: Creación de una base de datos *(Data Warehouse)* que integre los datos de diversas fuentes, utilizando métodos comoextracción estática, llamadas a API y web scraping.

Análisis de sentimientos: Aplicación de técnicas de procesamiento de *lenguaje natural (NLP)* para analizar el sentimiento de las reseñas y clasificarlas en positivas, negativas o neutrales.

Predicción de tendencias: Desarrollo de modelos de machine learning, supervisados o no supervisados, para predecir los rubros de los negocios que experimentarán crecimiento o declive en base a las reseñas.

Sistemas de recomendación: Implementación de un sistema de recomendación de negocios basado en las preferencias de los usuarios, que permita a los usuarios descubrir nuevos lugares en función de sus experiencias previas.

Análisis adicional: Cruzamiento de datos de reseñas con información como cotizaciones en bolsa, tendencias en redes sociales y medios de comunicación sobre comercios en expansión para obtener una visión más completa

En base a las reseñas de los usuarios, es posible identificar señales y patrones que pueden indicar el crecimiento o declive de un negocio.

KPI's

objetivo general:

- Determinar la satisfacción del cliente al utilizar el servicio de la plataforma Yelp o Google.
- Determinar la retención y abandono de los clientes.
- Determinar los rubros de negocios que más crecieron o decayeron.
- Predecir la localización más conveniente de nuevos locales de estos negocios evaluados.
- Implementar un sistema de recomendaciones para usuarios.

KPI	DESCRIPCION	FORMULA	FRECUENCIA
Promedio de calificacion	Tasa variación porcentual del promedio de calificación de estrellas del 3%	PCC= (promedio estrellas año anterior) - (promedio estrellas año actual) / (prromedio estrellas año actual) *100	Anual
Cierre de sucursales	Porcentaje de sucursales cerradas por rubro	CSC = (cantidad de sucursales cerradas) / (cantidad de sucursales totales por rubro) *100	Mensual
Satisfaccion del cliente	Satisfacción del cliente por reseña	(calificación de estrellas)	Semanal
Abandono del cliente al negocio	Reduccion del 5% la tasa de abadono global	TAG: (clientes perdidos)/ (total de clientes al inicio del ano)*100	Anual
Capacidad de retencion del cliente	Indice de tasa de retención de clientes	IRTC=((CE-CN)/CS)) *100 CE = El número total de clientes al final del periodo CN = El número total de nuevos clientes que ha adquirido durante el periodo CS = El número total de clientes al inicio de un periodo	Mensual
Abandono de cliente mensual	Tasa de abandono de cliente	TAC ((total de clientes al principio del mes - total de clientes al final del mes)/- total de clientes al principio del mes)*100	Mensual
Reduccion de perdida de cliente	Reducir los clientes perdidos en un 5%	CP (total de clientes al inicio del mes(año) - total de clientes al final mes(año))	Mensual

Los **KPIs** son herramientas que permiten medir la ***eficacia y la productividad*** de determinadas acciones, con el fin de saber si se están cumpliendo los objetivos establecidos. En definitiva, son un elemento clave a la hora de optimizar tu sistema de planificación.
Estas son las métricas que tienen un mayor impacto en el negocio, por lo tanto, representa un impacto positivo para la empresa.

Opiniones de clientes en línea:

Esta puede parecer obvia, pero es una de las mediciones más importantes de la satisfacción del cliente. Las plataformas de reseñas en línea más comunes, como Yelp o Google, emplean sus propias métricas para determinar si su empresa cumple con las expectativas de los clientes. Normalmente, se califican con estrellas, siendo 5 estrellas la mejor calificación y 1 estrella la peor.
Los estudios demuestran que la mayoría de la gente lee las reseñas antes de visitar un negocio, y entre los jóvenes de 18 a 34 años, el 91% de los encuestados confía en las reseñas tanto como en las recomendaciones personales. Por este motivo, animar a sus clientes a dejar reseñas después de comprar su producto o servicio es fundamental para atraer a más clientes.

Tasa de abandono:

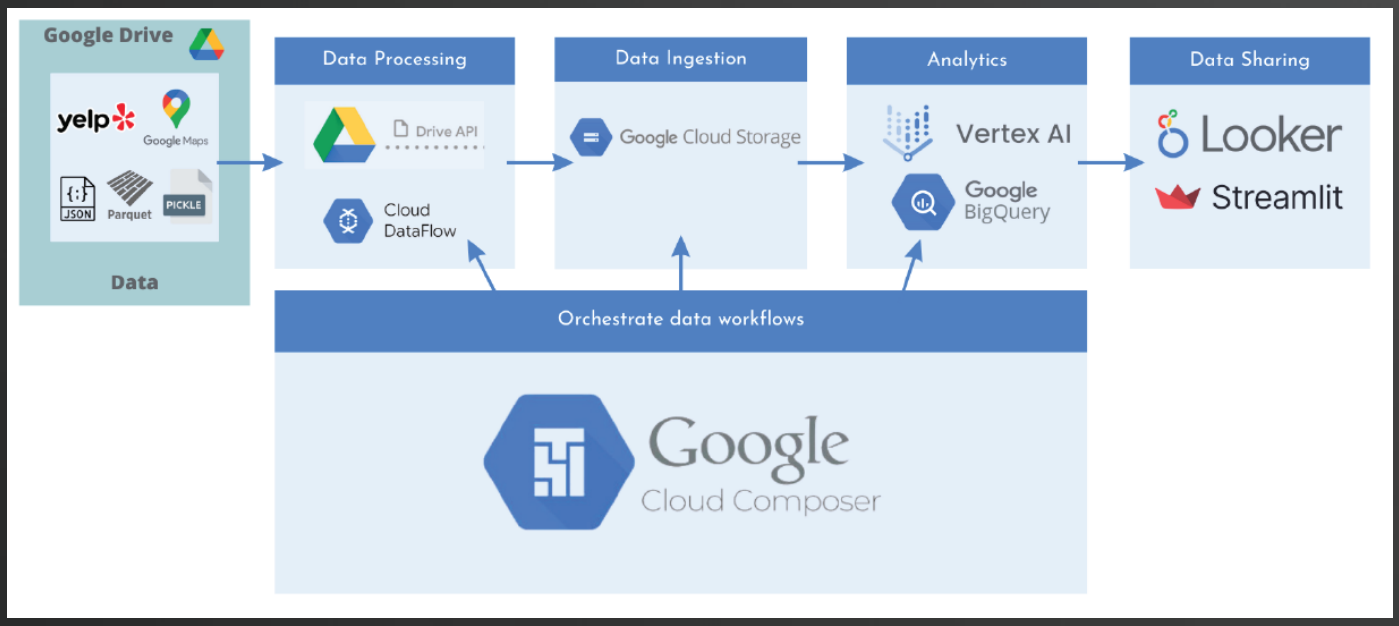
Es fundamental porque puede ayudar a las empresas a entender por qué los clientes se van e incluso indicar las épocas del año en las que la tasa de abandono es más alta. Los equipos de marketing y ventas pueden entonces tomar decisiones significativas en relación con nuevas promocionevs o desarrollar estrategias de activación de clientes basadas en la información. Puede ayudarle a cuidar mejor a sus clientes actuales y convertirlos en valiosos promotores de la marca.

Tasa de retención de clientes:

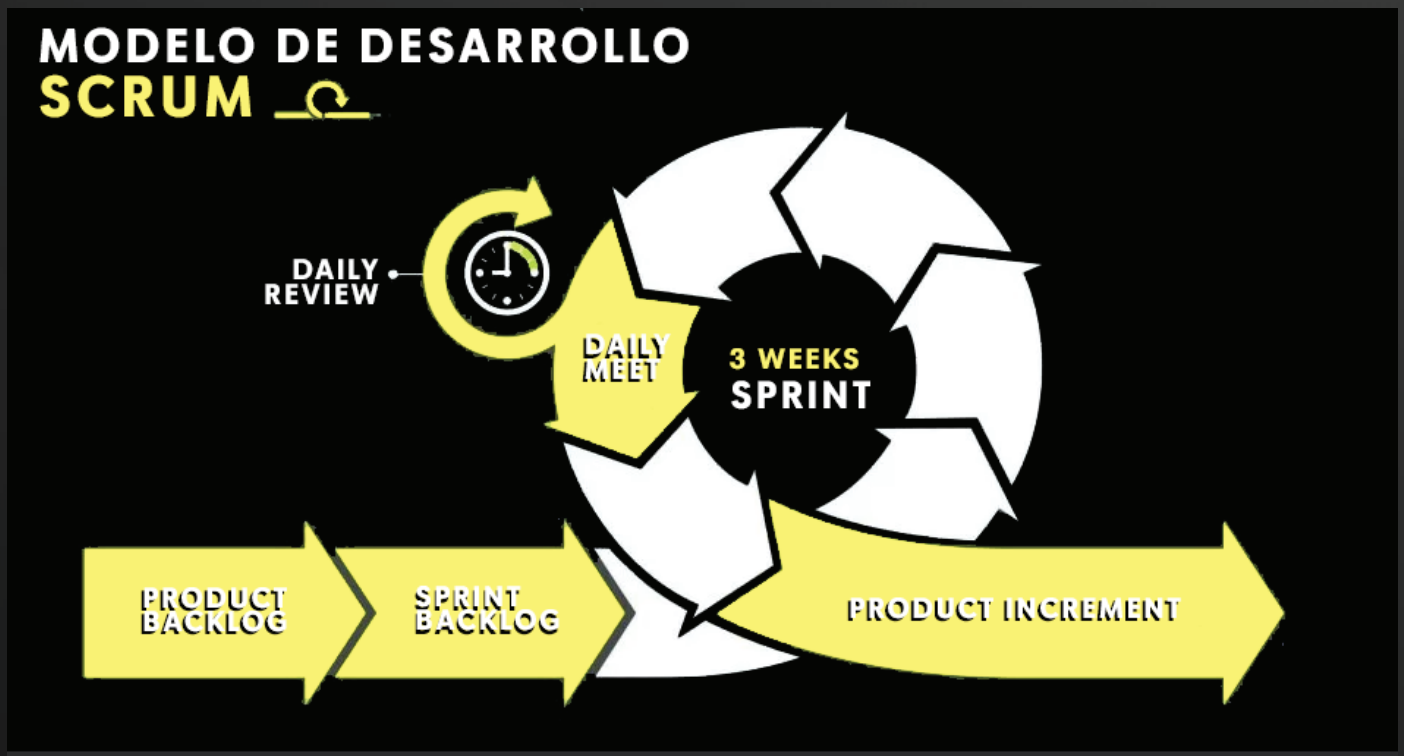
La retención de clientes es la capacidad de hacer que sus clientes vuelvan durante un periodo de tiempo. Retener a un cliente indica que su producto, servicio o marca es lo suficientemente satisfactorio como para que se quede con usted en lugar de cambiarse a un competidor.

PROPUESTA

stack tecnologico:



metodologia y equipo de trabajo :



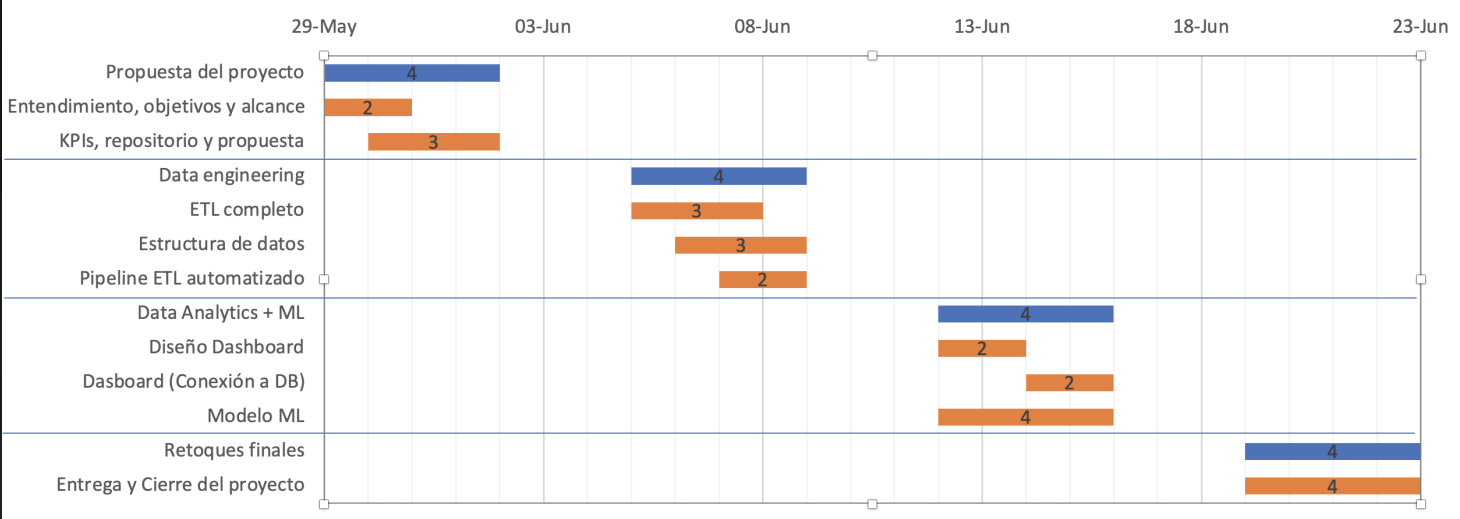
SCRUM ROLES:

PRODUCT OWNER:
CAROLINA VILLAGRA

SCRUM MASTER:
JULIAN MEDIAVILLA

TEAM MEMBERS:
CLARITZO PEREZ /data_analyst
PAULA PALLARES /functional_analysis
BENJAMIN ZAMBELLI /data_engineer
BEDER RIVERA /data_engineer
GONZALO SCHWERDT /ml_engineer

cronograma grantt:



Nombre actividad	Inicio	Duración	Fin
Propuesta del proyecto	29-May	4	02-Jun
Entendimiento, objetivos y alcance	29-May	2	31-May
KPIs, repositorio y propuesta	30-May	3	02-Jun
Data engineering	05-Jun	4	09-Jun
ETL completo	05-Jun	3	08-Jun
Estructura de datos	06-Jun	3	09-Jun
Pipeline ETL automatizado	07-Jun	2	09-Jun
Data Analytics + ML	12-Jun	4	16-Jun
Diseño Dashboard	12-Jun	2	14-Jun
Dasboard (Conexión a DB)	14-Jun	2	16-Jun
Modelo ML	12-Jun	4	16-Jun
Retoques finales	19-Jun	4	23-Jun
Entrega y Cierre del proyecto	19-Jun	4	23-Jun

ANALISIS PRELIMINAR DE DATOS

El **análisis preliminar de EDA** (Exploratory Data Analysis) es un proceso inicial en el análisis que tiene como objetivo explorar y comprender los datos antes de aplicar técnicas más avanzadas. Durante esta etapa, se examinan los datos en bruto para identificar patrones, tendencias, distribuciones, relaciones y posibles valores atípicos o errores. Algunas de las técnicas comunes utilizadas en el análisis preliminar de EDA son la visualizaciones, cálculos estadísticos descriptivos, gráficos, histogramas, gráficos de dispersión y análisis de correlación. El análisis preliminar ayuda a los científicos a obtener una comprensión inicial de los datos y tomar decisiones informadas sobre los pasos subsiguientes en el análisis.

El **análisis de calidad de datos** se refiere a evaluar y garantizar la calidad de los datos utilizados en un proyecto. Esto implica identificar problemas o errores en los datos que puedan afectar la precisión y confiabilidad de los resultados. Algunos aspectos comunes de la calidad de los mismos que se analizan incluyen la completitud, la precisión (exactitud de los datos), la consistencia (coherencia de los datos) y la validez (conformidad con reglas o restricciones). Además, se pueden verificar otros aspectos como la consistencia temporal, la consistencia de formatos y la detección de valores atípicos. El análisis de calidad de datos busca corregir o eliminar cualquier problema identificado antes de proceder con el análisis e interpretación de los datos.

En resumen, el **análisis preliminar de EDA** *se centra en explorar y comprender los datos, mientras que el* **análisis de calidad de datos** *se centra en garantizar que los datos utilizados sean confiables y estén libres de errores. Ambos son pasos importantes en el proceso de análisis de datos en la ciencia de datos y contribuyen a resultados más precisos y significativos.*

Google Maps dataset metadata-sitios

Observaciones:

Hay 3.025.011 filas y 15 columnas.

Número de valores nulos:

"MISC": 690.834

"dirección": 80.511

"categoría": 17.419

"descripción": 2.770.722

"horario": 787.405

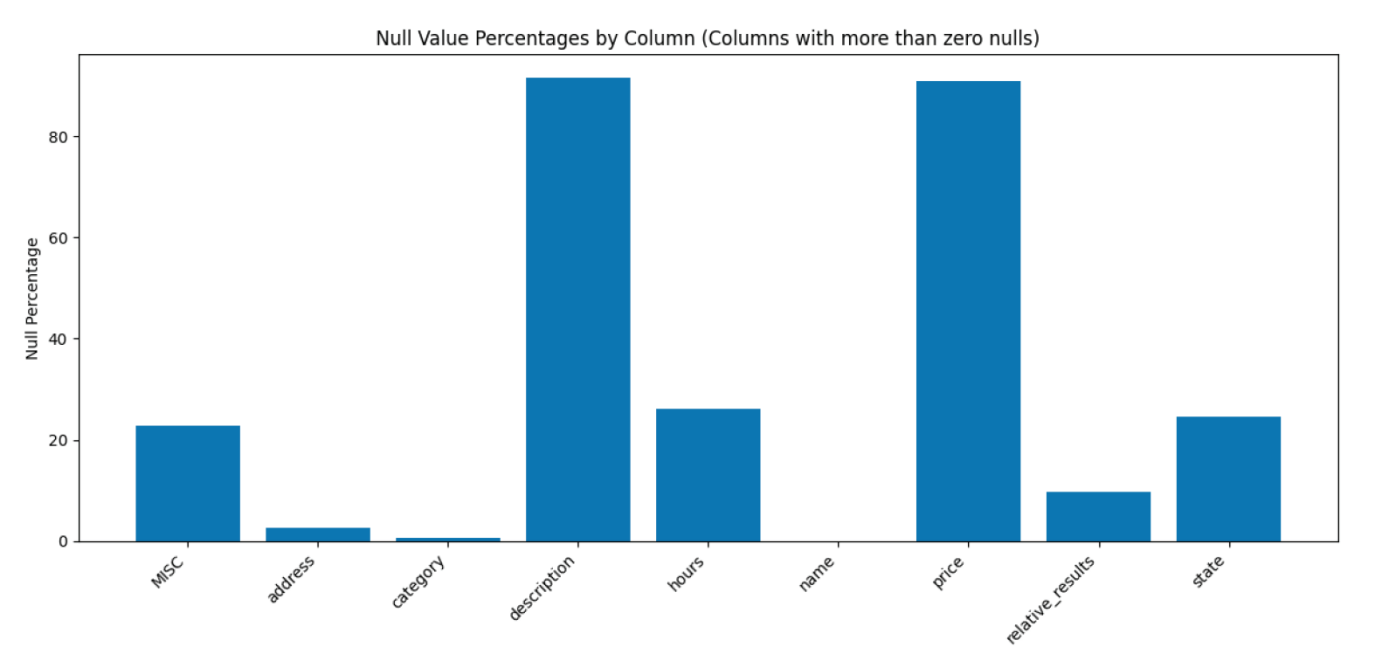
"nombre": 37

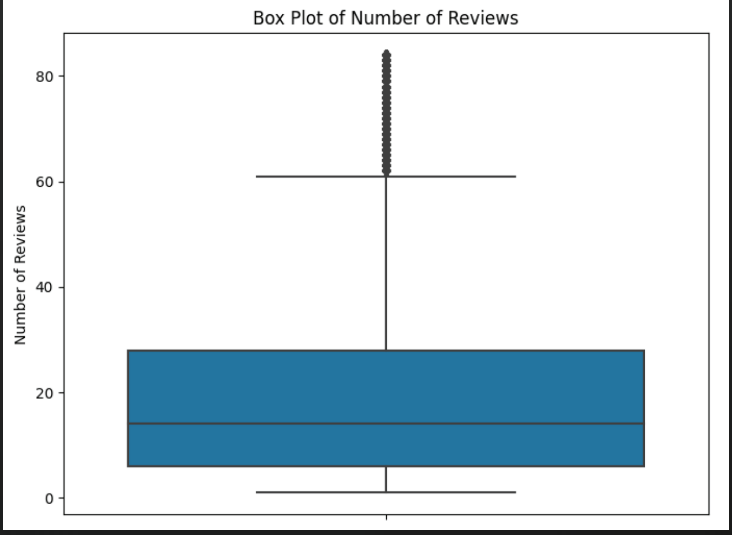
"precio": 2.749.808

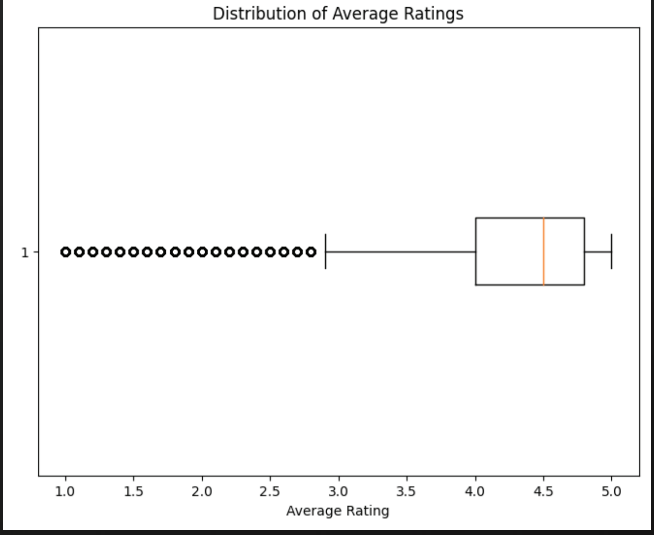
"resultados_relativos": 295.058

"estado": 746.455

Parece haber 26.573 duplicados







Google Maps

reviews-estados

Observaciones:

Hay 89.946.359 filas y 9 columnas.

Hay valores nulos en las columnas:

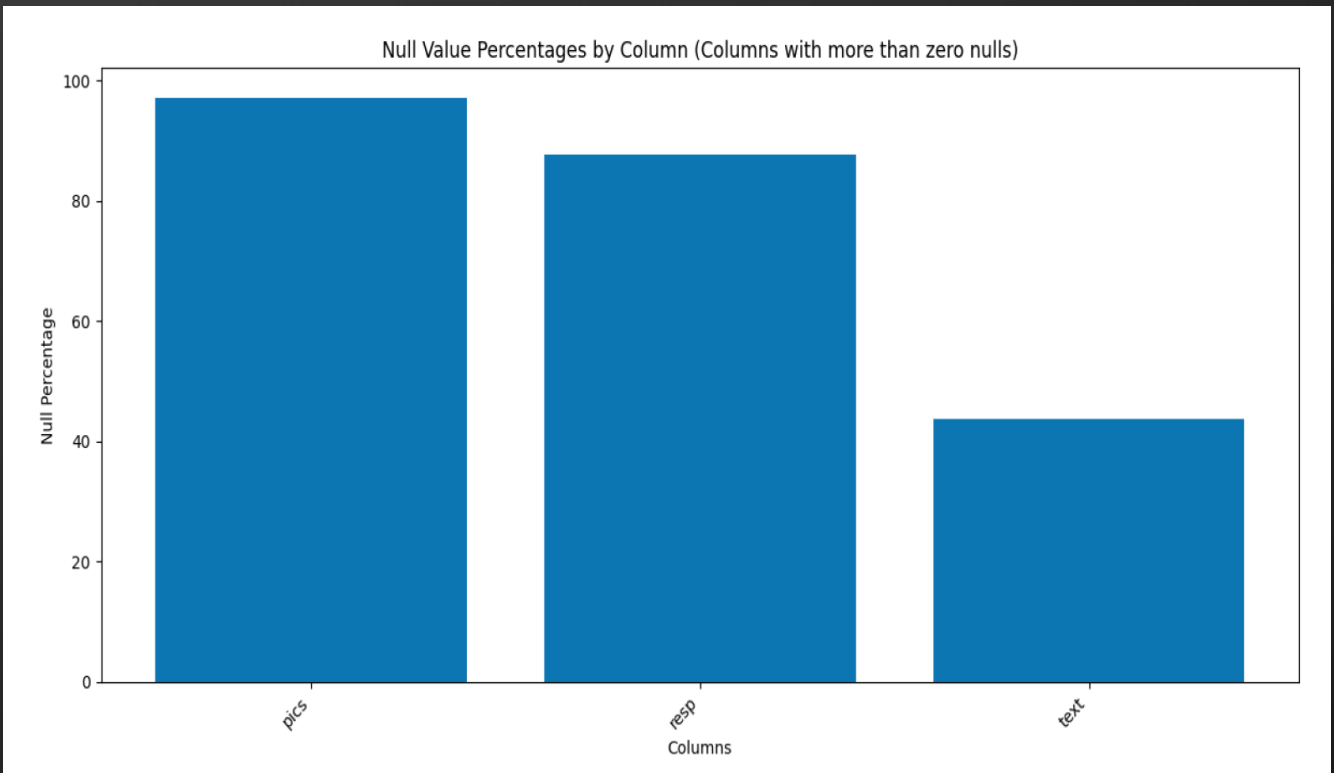
"imágenes" (87.450.680)

"respuesta" (78.917.249)

"texto" (39.307.744)

Parece haber 1.386.843 duplicados

La columna "tiempo" está en formato de sello de tiempo Unix, por lo que debemos convertirla a tipo de fecha para el análisis.



Yelp! dataset business.pkl

Observaciones:

Hay 150.346 filas y 28 (14) columnas.

¡Las columnas se repiten dos veces!

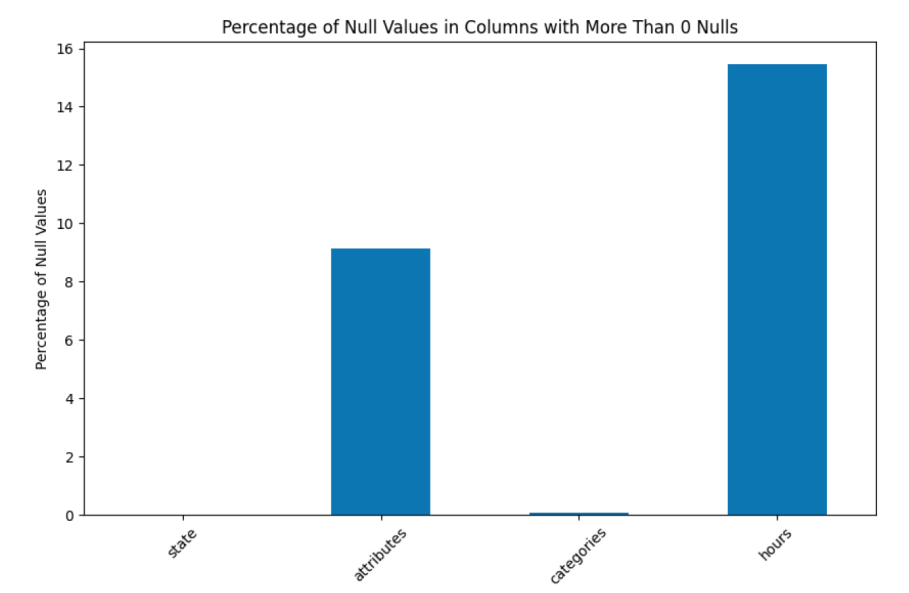
Aparentemente no hay duplicados

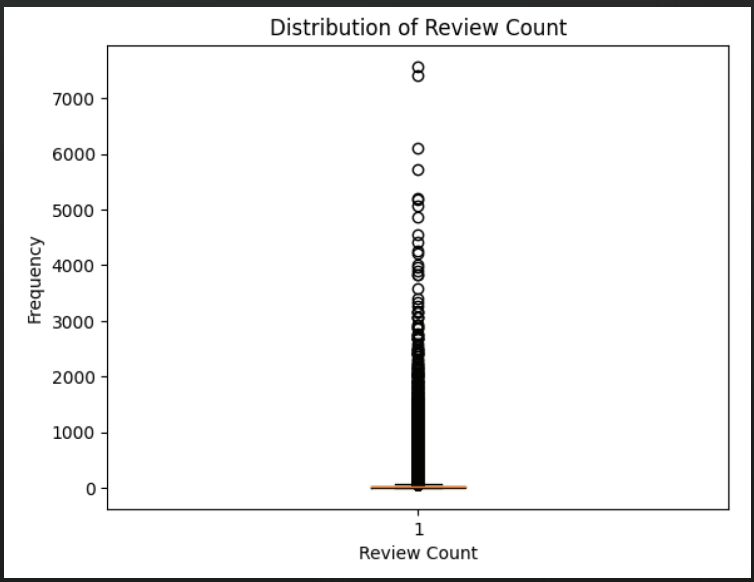
Todo es de tipo objeto. Las columnas 'estrellas', 'latitud' y 'longitud' deberían ser de tipo flotante, las columnas 'cantidad_de_reseñas' y 'está_abierto' deberían ser de tipo entero

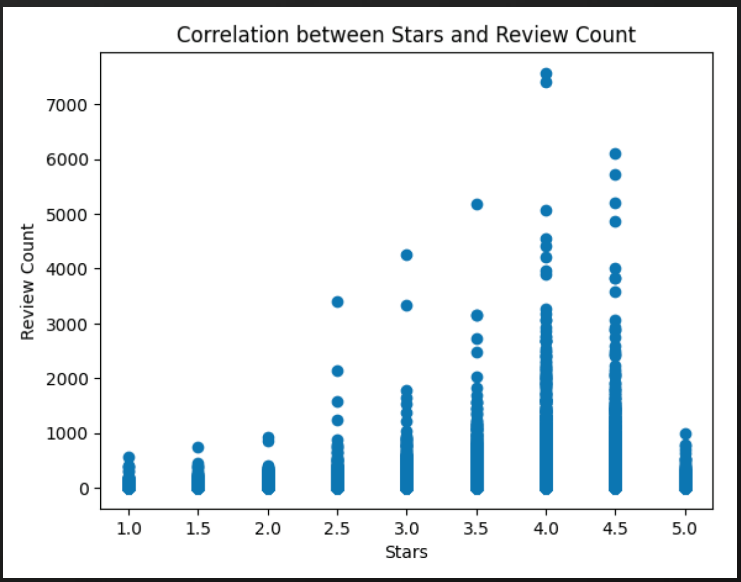
Hay valores nulos en las columnas: estado (3), atributos (13744), categorías (103), horarios (23223)

Hay algunos lugares con muchas más reseñas (posibles valores atípicos).

Parece haber muchas reseñas entre 2.5 y 4.5 estrellas y menos entre 1 y 2, y específicamente 5.







review.json

Observaciones:

Hay 6.990.280 filas y 9 columnas.

No parece haber valores nulos o duplicados.

La columna de fecha es de tipo cadena de texto.

user.parquet

Observaciones:

Hay 2.105.597 filas y 22 columnas.

No parece haber valores nulos.

Parece haber 117.700 valores duplicados.

GRACIAS 🚀

DATABRIDGE
SMART SOLUTIONS