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 usuarios respecto a diversos negocios, incluyendo restaurantes, hoteles, esteticas 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 cuidado personal y la estetica en el mercado estadounidense.

El rubro de belleza abarca una amplia gama de servicios y establecimientos relacionados con el cuidado personal y la estética. Algunos ejemplos de negocios dentro de este rubro son los salones de belleza, spas, peluquerías, barberías, salones de uñas, centros de estética, salones de masajes y tiendas de productos de belleza.

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 determinar las ubicaciones más adecuadas para establecer nuevos locales comerciales y descubrir oportunidades de inversión investigando aspectos como el crecimiento del mercado, la demanda de servicios de belleza, la competencia existente y las tendencias emergentes.

Con base en el análisis realizado, generaremos recomendaciones claras y fundamentadas para el **inversor**. Estas recomendaciones mostrarán las oportunidades de inversión más atractivas en el rubro de belleza, destacando los aspectos clave que respaldan la viabilidad y el potencial de crecimiento de cada oportunidad.

Aunque nos enfocaremos principalmente en el sector de estetica, la metodología puede aplicarse a otros tipos de comercios.

OBJETIVO

El objetivo principal del proyecto es brindar a nuestro cliente: inversor de la industria estetica latinoamericana una vision general del mercado estadounidense con el fin de que tome las decisiones mas informadas e inteligentes para incorporarse como competidor en dicho mercado. Gracias a un **análisis exhaustivo de la opinión de los usuarios en Yelp y Google Maps** podremos **identificar tendencias, predecir el crecimiento o decaimiento de rubros comerciales y tomar decisiones estratégicas informadas para mejorar decisiones de gestion e inversion de negocios**.

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: Basado en los hallazgos obtenidos, genera recomendaciones accionables para mejorar la experiencia del cliente, optimizar las estrategias de marketing, identificar oportunidades de negocio, etc.

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

KPI's

KPI	DESCRIPCION	FORMULA	FRECUENCIA
Volumen de reviews **(popularidad)**	aumentar en un 5% la cantidad de reviews respecto al mes anterior	((cant. reviews del mes actual - cant. reviews del mes anterior) / cant. reviews del mes anterior) * 100	Mensual
Proporcion positiva de reviews **(baja negatividad)**	La de reviews positivas debe mantenerse en 5:1 sobre las negativas. Considerando Review positiva > 4	cant. reviews positivas / cant. reviews negativas	Mensual
Satisfaccion del cliente **(mejora del servicio)**	Aumentar a 3% el promedio de calificaciones	((promedio calificacion actual - promedio calificacion anterior) / promedio calificacion anterior) * 100	Mensual
Aumento de catidad de locales por estado **(aumento de demanda)**	Aumentar en un 2% la cantidad de negocios por Estado	((cantidad negocios actual - cantidad de negocios anterior) / cantidad de negocios anterior) * 100	Mensual
Rubro belleza vs rubro restaurant **(competencia)**	Comparar el promedio de calificaciones del rubro belleza contra el rubro restaurant	((Promedio Calificaciones Belleza - Promedio Calificaciones Restaurant) / Promedio Calificaciones Belleza) * 100	Mensual

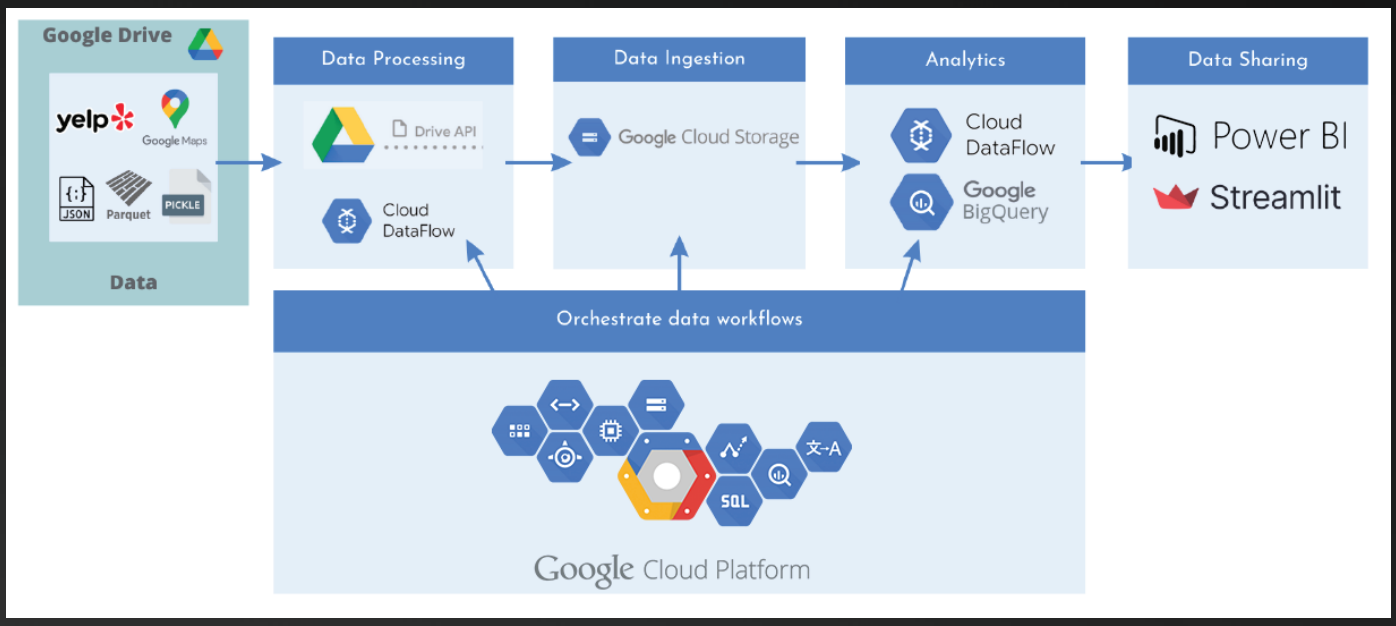
Opiniones de clientes en línea:

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.

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.

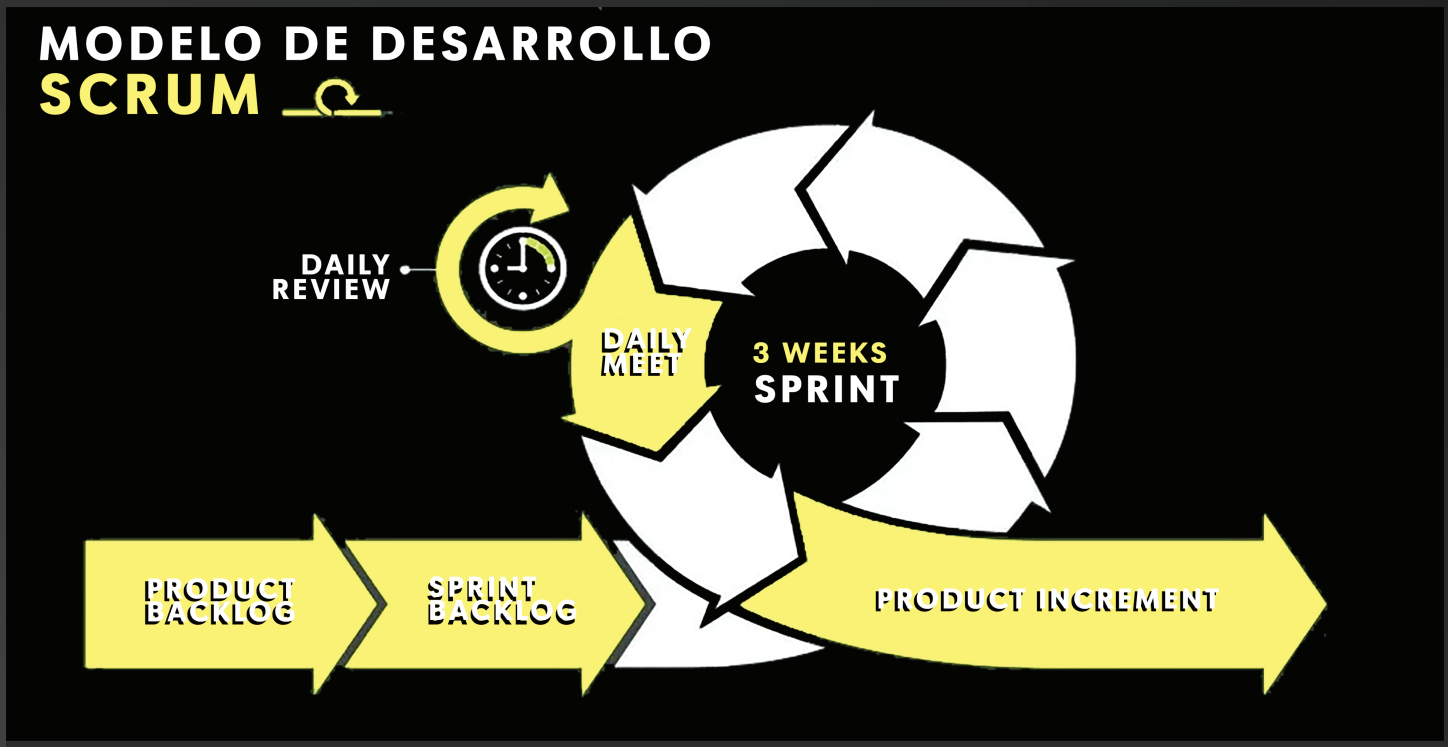
PROPUESTA

stack tecnologico:



metodologia y equipo de trabajo :

MODELO DE DESARROLLO SCRUM



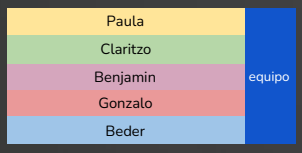
SCRUM ROLES:

PRODUCT OWNER:
CAROLINA VILLARRAGA

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:



		29/05	30/05	31/05	01/06	02/06	05/06	06/06	07/06	08/06	09/06	12/06	13/06	14/06	15/06	16/06	19/06	20/06	21/06
SEMANA 1	Propuesta del proyecto																		
	Planteo objetivos y alcance																		
	KPI's																		
	Armado de GitHub																		
	Stack tecnologico																		
	Propuesta de solucion																		
	Diagrama de gantt																		
	Analisis preliminar de datos																		
	Documentacion																		
SEMANA 2	Diseno de modelo de ER																		
	Pipeline DW																		
	Automatizacion DW + ETL																		
	Diccionario de datos																		
	EDA																		
	Documentacion																		
SEMANA 3	Conexion BigQuery con PowerBi																		
	Conexion BigQuery con python																		
	Dashboards																		
	KPI's																		
	Modelado ML																		
	Analisis de sentimiento																		
	Documentacion (feature engineer)																		
	Documentacion (analisis)																		
SEMANA 4	Documentacion a GitHub																		
	Storytelling																		
	Presentaciones (slides, videos)																		
	Analisis de funcionalidad																		
	Test de ML funcionando																		

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

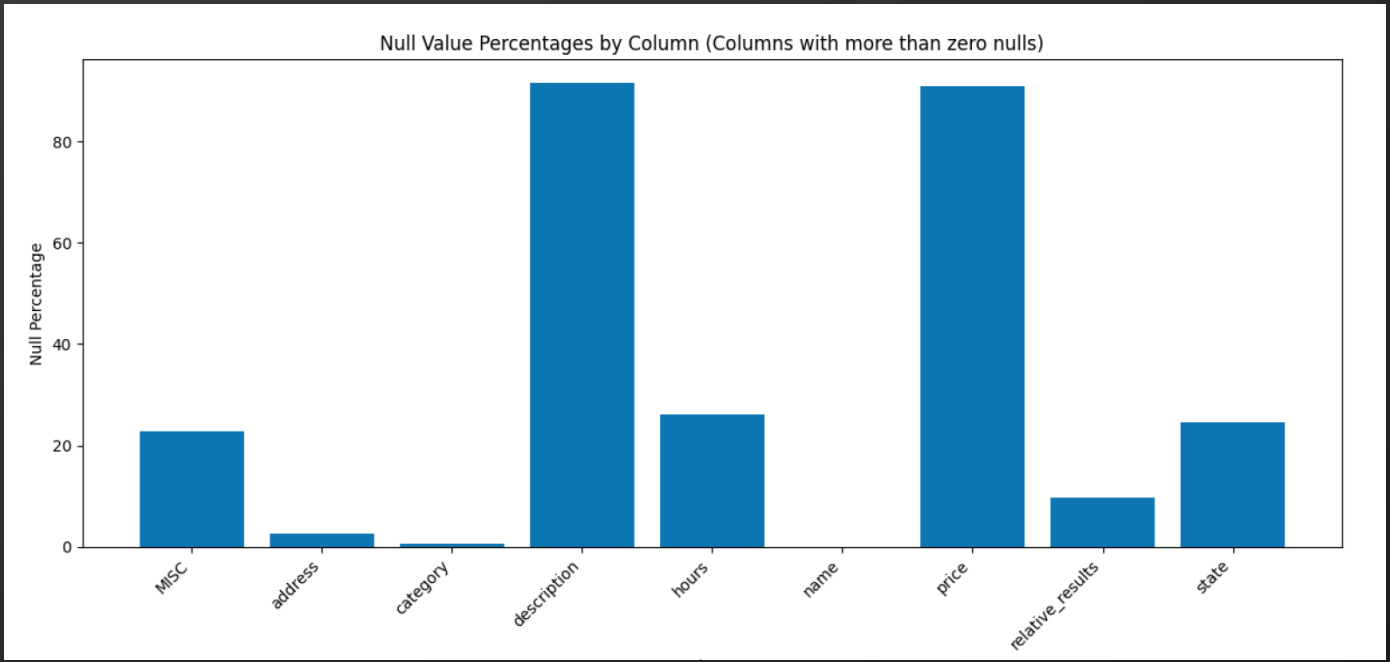
"nombre": 37

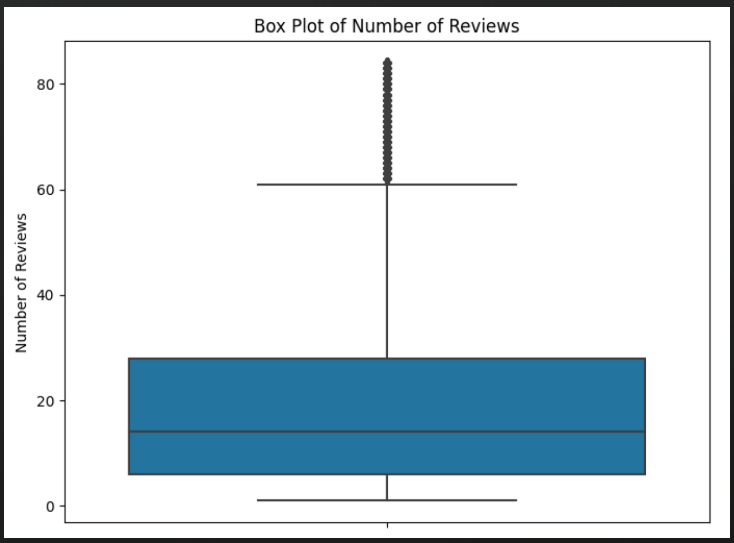
"precio": 2.749.808

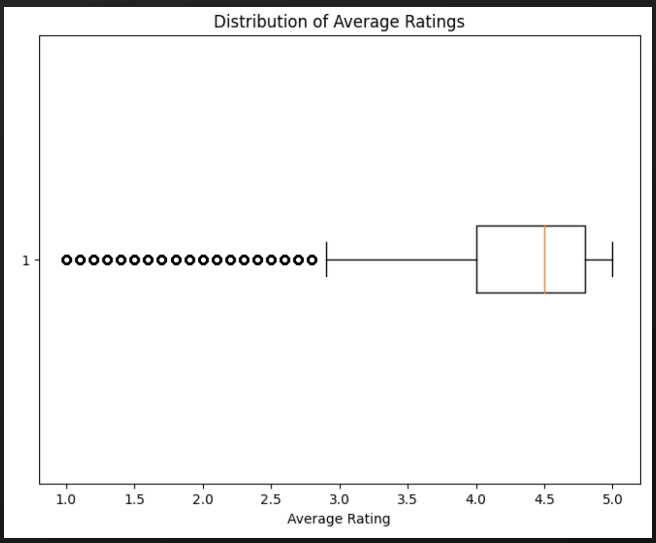
"resultados_relativos": 295.058

"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:

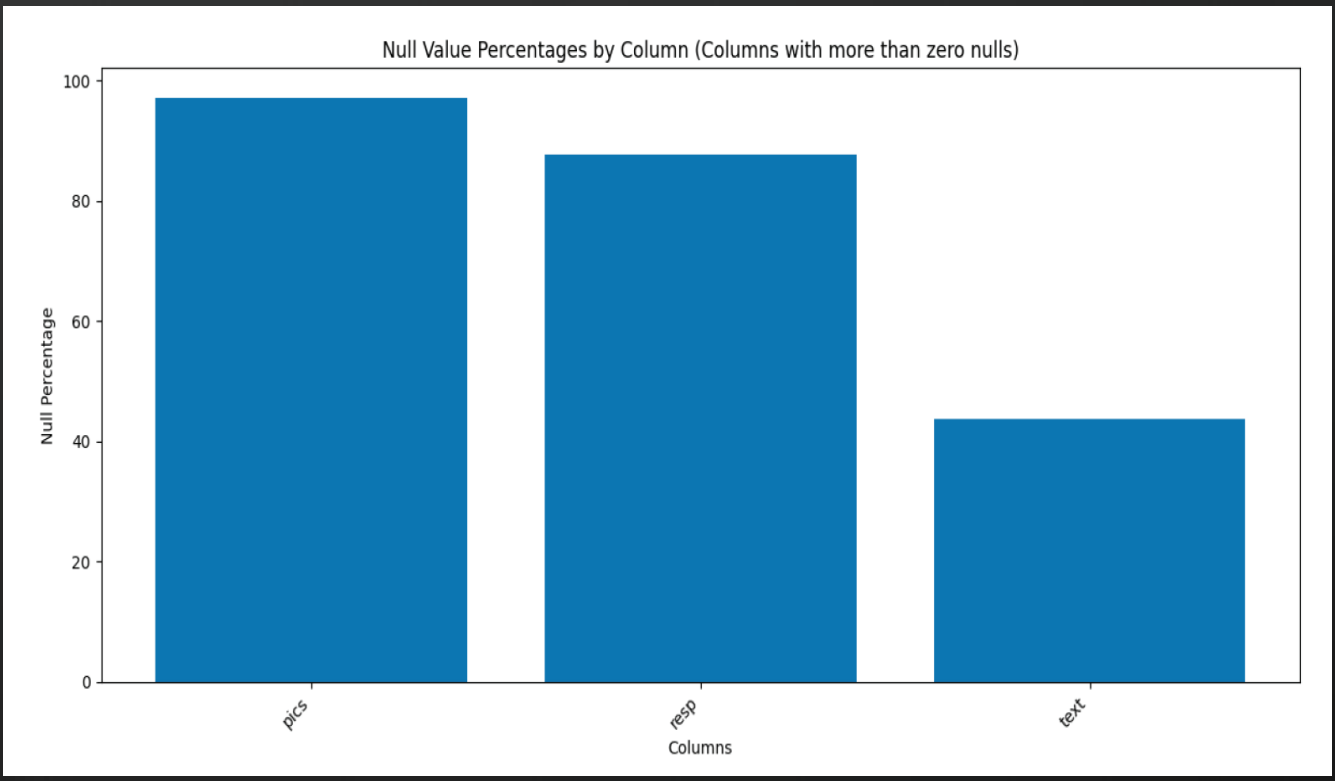
"imágenes" (87.450.680)

"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!

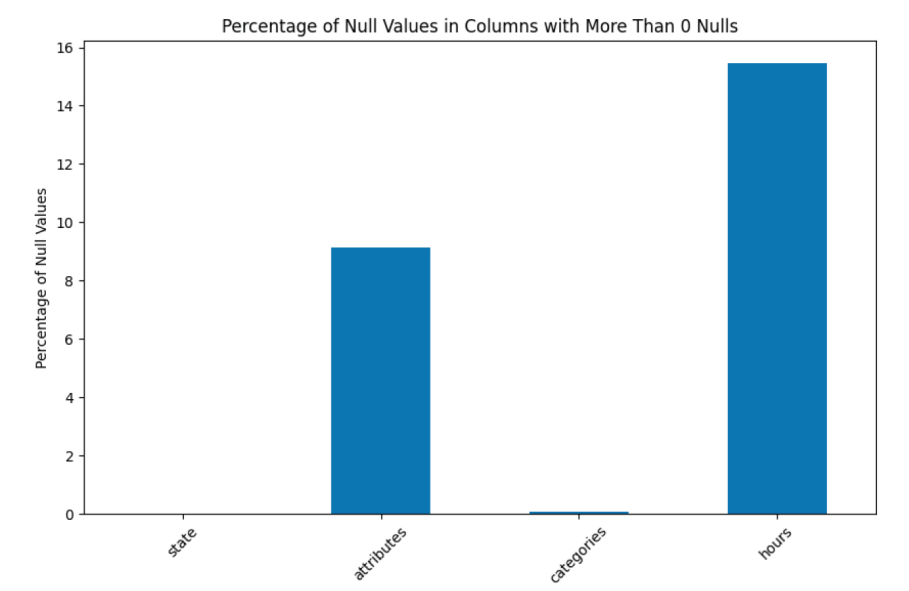
Aparentemente no hay duplicados

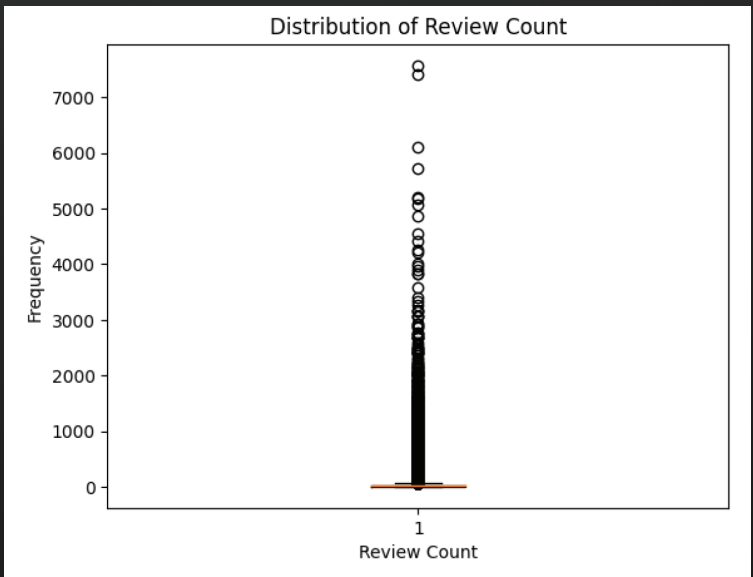
Todo es de tipo objeto. Las columnas 'estrellas', 'latitud' y 'longitud' deberían ser de tipo flotante, las columnas 'cantidad_de_reseñas' y 'está_abierto' deberían ser de tipo entero

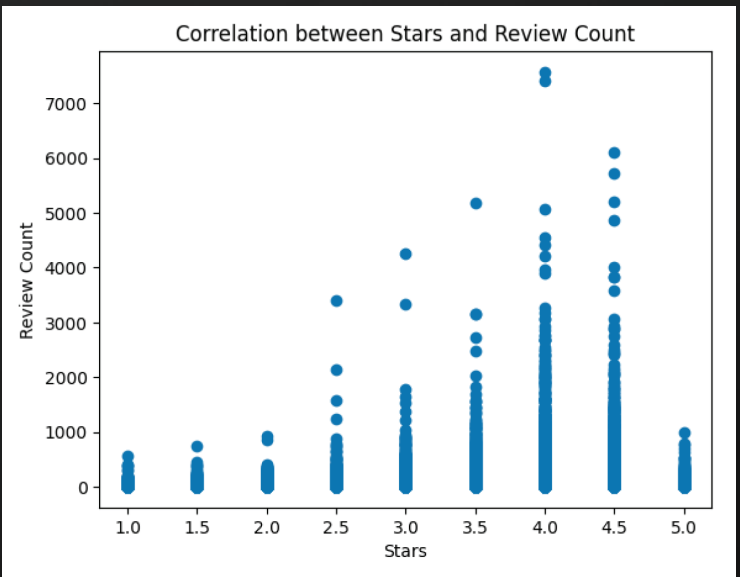
Hay valores nulos en las columnas: estado (3), atributos (13744), categorías (103), horarios (23223)

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