



DOCUMENTACION

El presente documento establece la arquitectura y el modelo de los datos que usará nuestra aplicación final.

Se ha decidido usar la plataforma de **Google Cloud** para almacenar y gestionar nuestros datos, hemos elegido Google BigQuery como nuestro sistema de almacenamiento en la nube. Esta elección se basa en varias ventajas clave que BigQuery ofrece, como su escalabilidad, rendimiento y facilidad de uso. Además, BigQuery es un servicio totalmente administrado, lo que significa que nos libera de la carga de administrar la infraestructura subyacente, permitiéndonos centrarnos en el análisis de datos y la toma de decisiones.

En primer lugar se describe la **plataforma tecnológica** seleccionada, así como el flujo de los datos desde su almacenamiento original hasta su preparación en la solución propuesta.

En segundo lugar se describen los **procesos de automatización** que dan soporte al ciclo de vida de los datos.

Por ultimo se detalla la **estructura de los datos y las relaciones** entre los diferentes datos.



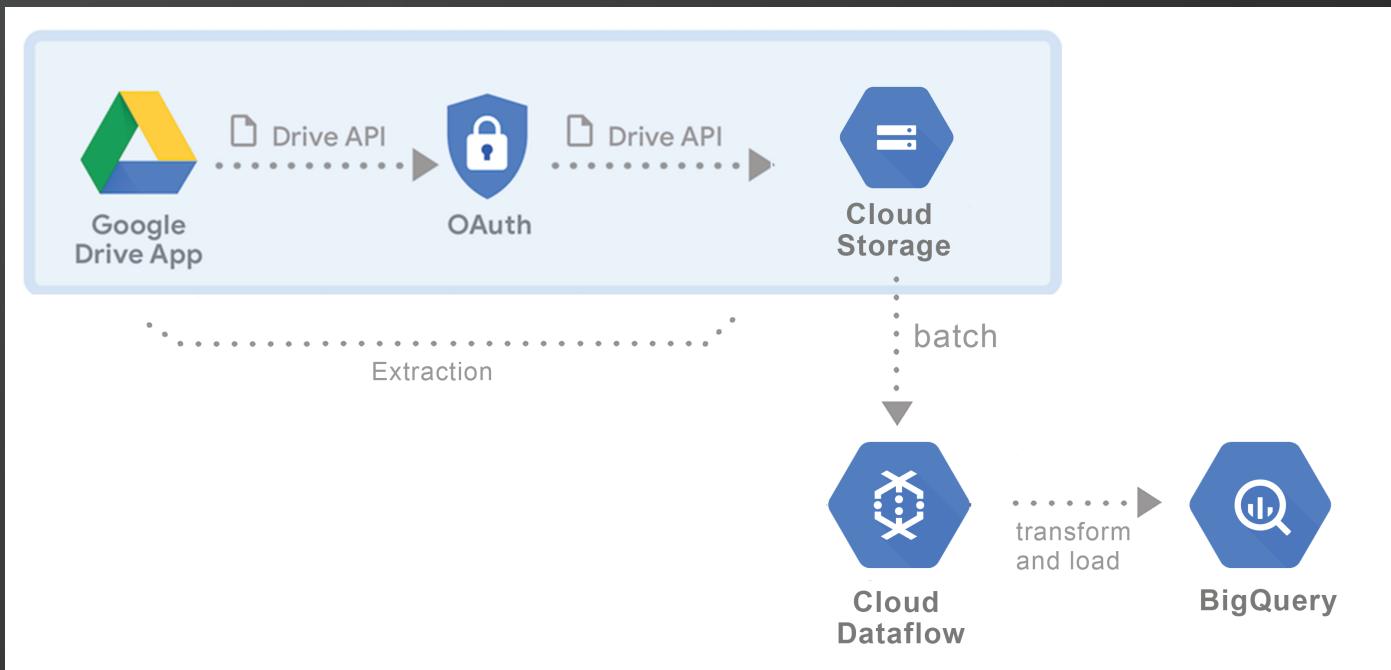
Google Cloud Platform

Google Cloud Platform (GCP) es una plataforma de servicios en la nube proporcionada por Google que ofrece una amplia gama de herramientas y servicios incluyendo almacenamiento, procesamiento, análisis, inteligencia artificial, aprendizaje automático, bases de datos y más. Estos servicios están diseñados para brindar una infraestructura confiable y escalable para ejecutar aplicaciones y almacenar datos en la nube.

En nuestro proyecto, hemos decidido utilizar Google Cloud Platform por varias razones. En primer lugar, GCP nos proporciona una infraestructura altamente confiable. Google invierte significativamente en la infraestructura subyacente, asegurando una alta disponibilidad y protección de datos. Esto significa que podemos confiar en que nuestros datos y servicios estarán disponibles y seguros en todo momento. Además, GCP ofrece una escalabilidad sobresaliente. Podemos aumentar o disminuir fácilmente nuestros recursos según sea necesario, lo que nos permite manejar grandes volúmenes de datos y cargas de trabajo variables. Esto es especialmente importante en nuestro proyecto, donde trabajamos con datos grandes y realizamos operaciones de procesamiento intensivo.

Otro beneficio clave de GCP es su amplia gama de servicios y herramientas. Tenemos acceso a servicios como BigQuery, Cloud Storage, Dataflow, Composer y muchos más, que nos permiten realizar tareas como almacenar, procesar, analizar y orquestar datos de manera eficiente. Esto nos ayuda a construir soluciones complejas y escalables sin tener que preocuparnos por la infraestructura subyacente.

↗ FLUJO DE TRABAJO



El proceso completo comienza configurando **la API de Drive en Google Cloud** y creando un servicio con los permisos necesarios. Luego, en el script de ETL, nos autenticamos utilizando las credenciales obtenidas para acceder a la API de Drive.

A continuación, identificamos los archivos específicos en Google Drive que deseamos extraer y utilizamos la lógica implementada para recuperar esos archivos utilizando la API de Drive.

Una vez que hemos extraído los datos, configuramos **un bucket de Google Cloud Storage (GCS)** en nuestro proyecto.

Luego, creamos un **pipeline de Dataflow utilizando Apache Beam** para procesar los datos desde GCS. Definimos las transformaciones necesarias, *como limpiar, filtrar o agregar los datos, e implementamos la lógica de transformación* utilizando el modelo de programación de Apache Beam.

Después de procesar los datos, configuramos un conjunto de datos en **BigQuery** para almacenar los datos procesados. Definimos el esquema para las tablas de destino en BigQuery.

Utilizamos herramientas como Google Cloud Monitoring para obtener métricas de Dataflow y BigQuery y configuramos alertas o notificaciones para eventos críticos o fallos.

Finalmente, **automatizamos el proceso de ETL** mediante la programación, utilizando herramientas como Cloud Scheduler para ejecutar el pipeline de ETL en intervalos programados. Configuramos notificaciones o alertas para recibir información sobre ejecuciones exitosas o fallidas.

En resumen, el flujo de trabajo completo involucra la configuración de la API de Drive, la extracción de datos desde Google Drive, la carga de datos en GCS, el procesamiento de datos con Dataflow y Apache Beam, la carga de datos en BigQuery, el monitoreo y manejo de errores, y la programación y automatización del proceso. Esto nos permite tener datos limpios y ordenados en nuestro data warehouse de BigQuery, obteniendo así información valiosa y confiable para nuestro proyecto.



Gracias a este flujo de trabajo completo, que incluye la extracción, transformación y carga de datos en BigQuery, podemos obtener un data warehouse robusto y confiable. Esto nos brinda la base para realizar **análisis de negocio y crear dashboards interactivos** que nos permiten visualizar y explorar los datos de manera intuitiva.

Al tener los datos limpios y estructurados en BigQuery, **podemos realizar consultas y análisis avanzados** para extraer información significativa y obtener insights sobre nuestro negocio. *Estos análisis nos ayudan a identificar patrones, tendencias y relaciones entre los datos, lo que nos permite tomar decisiones más informadas y estratégicas.*

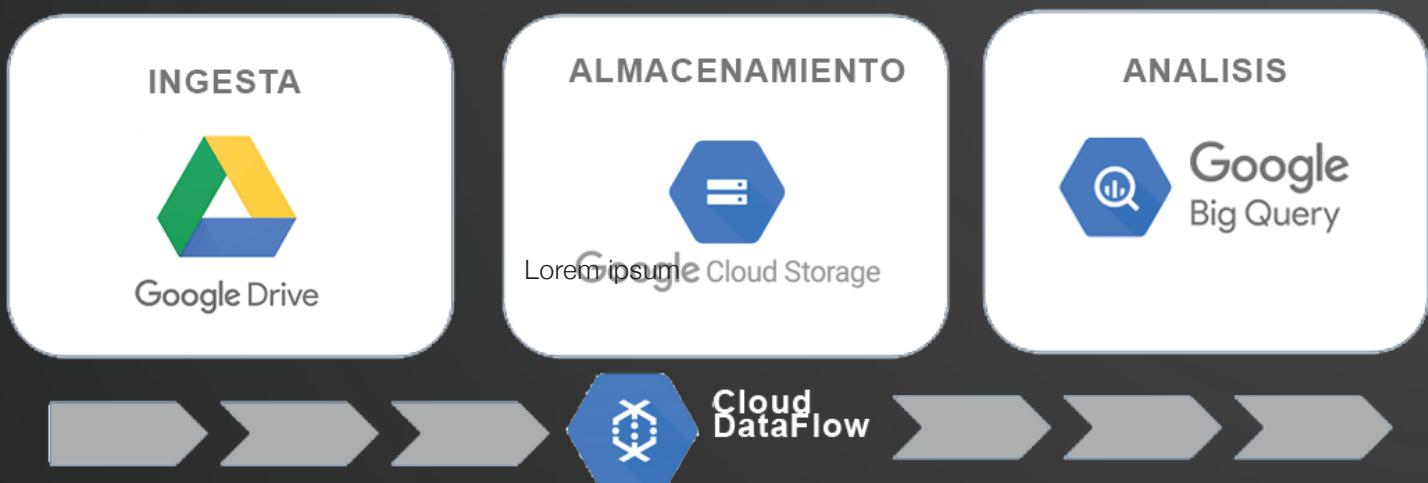
Además, al disponer de los datos en BigQuery, podemos aprovechar las capacidades de Google Cloud Platform para aplicar **técnicas de machine learning**. Podemos utilizar herramientas como Google Cloud AutoML o TensorFlow para construir modelos de machine learning y entrenarlos utilizando los datos almacenados en BigQuery. *Estos modelos pueden ayudarnos a predecir comportamientos, realizar recomendaciones personalizadas, optimizar procesos y tomar decisiones basadas en datos.*

En resumen, el flujo de trabajo completo, desde la ingestión de datos hasta la creación de un data warehouse en BigQuery, nos brinda la base para realizar análisis de negocio, crear dashboards interactivos y desarrollar modelos de machine learning. Esto nos permite aprovechar al máximo nuestros datos y utilizarlos como una ventaja competitiva en la toma de decisiones y la generación de valor en nuestro proyecto.

cronograma restante:

| 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) | | | | |
| SEMANA 4 | Documentacion (analisis) | | | | |
| | Documentacion a GitHub | | | | |
| | Storytelling | | | | |
| | Presentaciones (slides, videos) | | | | |
| | Analisis de funcionalidad | | | | |
| | Test de ML funcionando | | | | |

AUTOMATIZACION (ETL)



Al automatizar el proceso con Dataflow, podemos asegurarnos de que los datos estaran limpios y transformados de manera consistente y confiable antes de cargarlos en el datawarehouse.

Al eliminar duplicados, corregir inconsistencias y completar los datos faltantes, podemos obtener una visión clara y precisa de nuestro negocio.

Además, al tener datos limpios y ordenados, el proceso de análisis se vuelve más eficiente. Los analistas y profesionales de negocio pueden explorar y consultar los datos de manera más efectiva, ahorrando tiempo y esfuerzo en la búsqueda de información precisa. Esto nos permite generar informes más rápidos y fiables, así como realizar análisis en tiempo real para tomar decisiones ágiles.

Pasos para el ETL con Dataflow:

Definir el flujo de datos: esto implica determinar las fuentes de datos de entrada, las transformaciones que se aplicarán a los datos y la salida final hacia BigQuery.

Configurar el entorno: configurar el entorno de Dataflow en Google Cloud. Esto implica establecer las opciones de configuración, como el tamaño de los recursos, el número de trabajadores y la ubicación del almacenamiento.

Desarrollar las transformaciones: utilizar la programación basada en Apache Beam para desarrollar las transformaciones necesarias en el flujo de datos. Esto incluye tareas como limpiar los datos, filtrar registros no deseados, transformar formatos y calcular métricas.

Ejecutar el flujo de datos: una vez que se ha desarrollado el flujo de datos, ejecutamos las transformaciones en Dataflow. Esto iniciará la ejecución del proceso ETL, que leerá los datos de las fuentes, aplicará las transformaciones definidas y cargará los datos limpios y procesados en BigQuery.

Programar y automatizar: programamos la ejecución del flujo de datos en Dataflow utilizando herramientas como Cloud Scheduler o Cloud Functions. Esto nos permite establecer horarios regulares para la ejecución del proceso ETL y asegurar que se realice de manera automatizada.

Importancia de tener datos limpios y ordenados en el data warehouse:

Calidad de los datos: los datos limpios y ordenados garantizan la calidad de la información almacenada en el datawarehouse. Esto implica tener datos consistentes, precisos y completos, lo que proporciona una base sólida para el análisis y la toma de decisiones.

Eficiencia en el análisis: Los datos limpios y ordenados facilitan el proceso de análisis. Al eliminar registros duplicados, valores inconsistentes o datos faltantes, se obtiene una visión más clara y precisa de los datos, lo que facilita el análisis y la generación de informes.

Facilidad de mantenimiento: Mantener los datos limpios y ordenados en el data warehouse reduce la necesidad de realizar limpiezas y correcciones posteriores. Esto ahorra tiempo y esfuerzo en el mantenimiento continuo de los datos y permite centrarse en tareas de análisis y generación de información.



BigQuery es un almacén de datos y plataforma de análisis completamente administrado y sin servidor proporcionado por Google Cloud. Ofrece potentes capacidades de consulta SQL y puede manejar conjuntos de datos a gran escala de manera eficiente. Al cargar tus datos limpios desde GCS a BigQuery, puedes aprovechar las capacidades de consulta de BigQuery para realizar consultas analíticas complejas, generar ideas y crear visualizaciones. BigQuery también admite capacidades de aprendizaje automático a través de integraciones con Google Cloud Machine Learning Engine y AutoML.

Ventajas de BigQuery:

Escalabilidad: BigQuery es un servicio de almacenamiento y análisis de datos altamente escalable. Puede manejar grandes volúmenes de datos y realizar consultas rápidas incluso en conjuntos de datos masivos.

Rendimiento: BigQuery utiliza un motor de consulta distribuido y paralelo, lo que permite ejecutar consultas de manera eficiente y obtener resultados rápidos, incluso en tablas con miles de millones de filas.

Facilidad de uso: BigQuery utiliza un lenguaje de consulta similar a SQL, lo que facilita a los analistas y científicos de datos trabajar con él. También cuenta con una interfaz gráfica intuitiva y herramientas de desarrollo integradas.

Integración con el ecosistema de Google Cloud: BigQuery se integra de manera nativa con otros servicios de Google Cloud Platform, como Dataflow, Cloud Storage y Google Analytics. Esto permite una integración sin problemas en el flujo de datos y análisis.

Precio basado en el consumo: BigQuery utiliza un modelo de precios basado en el consumo, lo que significa que solo pagas por los recursos que realmente utilizas. Esto lo hace más flexible y rentable en comparación con las soluciones tradicionales de data warehousing.

Importancia de BigQuery para el proyecto:

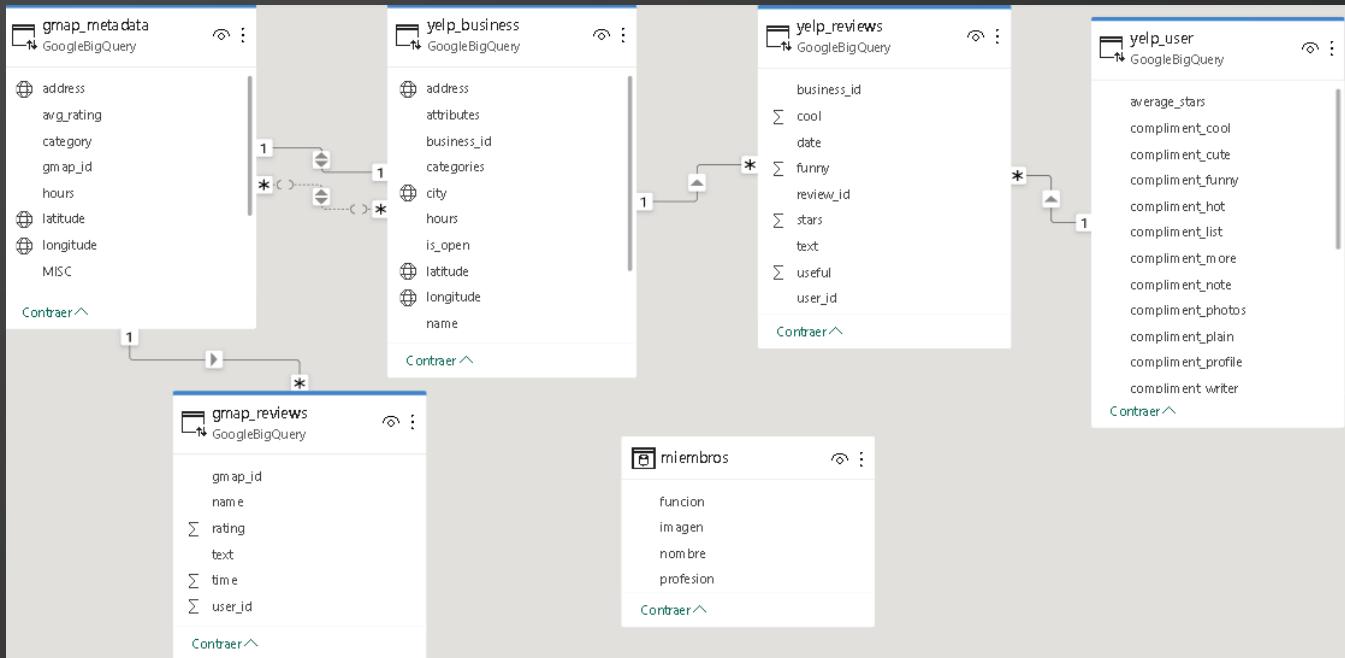
Almacenamiento centralizado: BigQuery actúa como el almacén central para los datos procesados en el pipeline ETL. Permite almacenar y consultar datos de manera eficiente, lo que facilita el análisis y la generación de información valiosa.

Escalabilidad y rendimiento: El proyecto implica el procesamiento de grandes volúmenes de datos, y BigQuery proporciona la escalabilidad y el rendimiento necesarios para manejar estos volúmenes y realizar consultas rápidas en ellos.

Integración con otras herramientas de análisis: BigQuery se integra con herramientas populares de análisis y visualización de datos, como Google Data Studio, lo que facilita el análisis y la generación de informes a partir de los datos almacenados.

En resumen, **BigQuery ofrece ventajas significativas en términos de escalabilidad, rendimiento, facilidad de uso y integración con otros servicios de Google Cloud Platform.** Su capacidad para almacenar y analizar grandes volúmenes de datos de manera eficiente lo convierte en una elección ideal para el proyecto, permitiendo un análisis de datos efectivo y una generación de información valiosa para la toma de decisiones.

MODEL DE ENTIDAD - RELACION



DICCIONARIO DE DATOS

Google Maps

metadata_sitios

La carpeta tiene 11 archivos ".JSON" donde se dispone la metadata del comercio.

Columnas:

- 'name': String, Nombre del local.
- 'address': String, Nombre del establecimiento, Número y nombre de la calle,Ciudad y Código postal
- 'gmap_id': String, ID único de la ubicación en Google Maps.
- 'description': Breve descripción del comercio.
- 'latitude': Float, Latitud
- 'longitude': Float, Longitud
- 'category': String, Categoría que clasifica el tipo de comercio.
- 'avg_rating': Float, Promedio del puntaje de las reseñas.
- 'num_of_reviews': Entero, Cantidad de reseñas.
- 'price': Entero, Precios.
- 'hours': String, dia/hora de atención
- 'MISC': Detalles adicionales:
 - 'Service options': String, Opciones de servicio.
 - 'Health & safety': String, Protocolos sanitarios.
 - 'Accessibility':String, Accesibilidad a nivel físico.
 - 'Planning': String, Tipo de plan semejante.
 - 'Payments': String, Tipo de pago disponible.
 - 'state': String, Estado de funcionamiento.
 - 'relative_results': String, códigos de ubicación geográfica.
 - 'url': String, Link de la búsqueda en Google Maps.

Google maps

review-estados

Reviews de los usuarios (51 carpetas, 1 por cada estado de USA, con varios archivos ".JSON" cada uno).

Columnas:

'user_id': String, ID único de usuario.
'name': String, Nombre y Apellido.
'time': Date, fecha de la reseña.
'rating': Entero, Número de calificación entre 1 y 5.
'text': String, Comentario de la reseña.
'pics': String, Lista de imágenes asociadas a la reseña vinculada cada una a una URL.
'resp':String, Respuesta proporcionada por el propietario del negocio o entidad relacionada con la reseña
'gmap_id': String, ID único de la ubicación en Google Maps.

Yelp

business.pkl

Información del comercio.

Columnas:

"business_id": string, 22 caracteres id del negocio
"name": string, nombre del negocio
"address": string, dirección completa del negocio
"city": string, ciudad
"state": string, código de 2 letras del Estado donde se ubica el negocio
"postal code": string, el código postal
"latitude": float, latitud
"longitude": float, longitud
"stars": float, rating en estrellas, redondeado a 0 o 0.5
"review_count": entero, número de reseñas
"is_open": entero, 0 si está cerrado, 1 si está abierto
"attributes": objeto, atributos del negocio como valores. Algunos valores de atributos también pueden ser objetos.
"categories" :lista de categorías de los negocios
"hours": objeto, dia/hora de atención

review.json

Contiene las reseñas completas, incluyendo el user_id que escribió el review y el business_id por el cual se escribe la reseña.

Columnas:

"review_id": string, 22 caracteres id de reseña
"user_id": string, 22 caracteres id único de usuario, refiere al usuario en user.json
"business_id": string, 22 caracteres id del negocio, refiere al negocio en business.json
"stars": entero, puntaje en estrellas de 1 al 5
"date": string, fecha formato YYYY-MM-DD
"text": string, la reseña en inglés
"useful": entero, números de votos como reseña útil
"funny": entero, número de votos como reseña graciosa
"cool": entero, número de votos como reseña cool.

user.parquet

Data del usuario incluyendo referencias a otros usuarios amigos y a toda la metadata asociada al usuario.

Columnas:

"user_id": string, 22 caracteres, id de usuario que refiere al usuario en user.json
"name": string, nombre del usuario
"review_count": entero, número de reseñas escritas
"yelping_since": string, fecha de creación del usuario en Yelp en formato YYYY-MM-DD
"friends": lista con los id de usuarios que son amigos de ese usuario
"useful": entero, número de votos marcados como útiles por el usuario
"funny": entero, número de votos marcados como graciosos por el usuario
"cool": entero, número de votos marcados como cool por el usuario
"fans": entero, número de fans que tiene el usuario
"elite": lista de enteros, años en los que el usuario fue miembro elite
"average_stars": float, promedio del valor de las reseñas
"compliment_hot": entero, total de cumplidos 'hot' recibidos por el usuario
"compliment_more": entero, total de cumplidos varios recibidos por el usuario
"compliment_profile": entero, total de cumplidos por el perfil recibidos por el usuario.

checkin.json

Registros en el negocio.

Columnas:

"business_id": string, 22 caracteres id del negocio
"date": string que es una lista de fechas separados por coma, en formato YYYY-MM-DD
HH:MM:SS

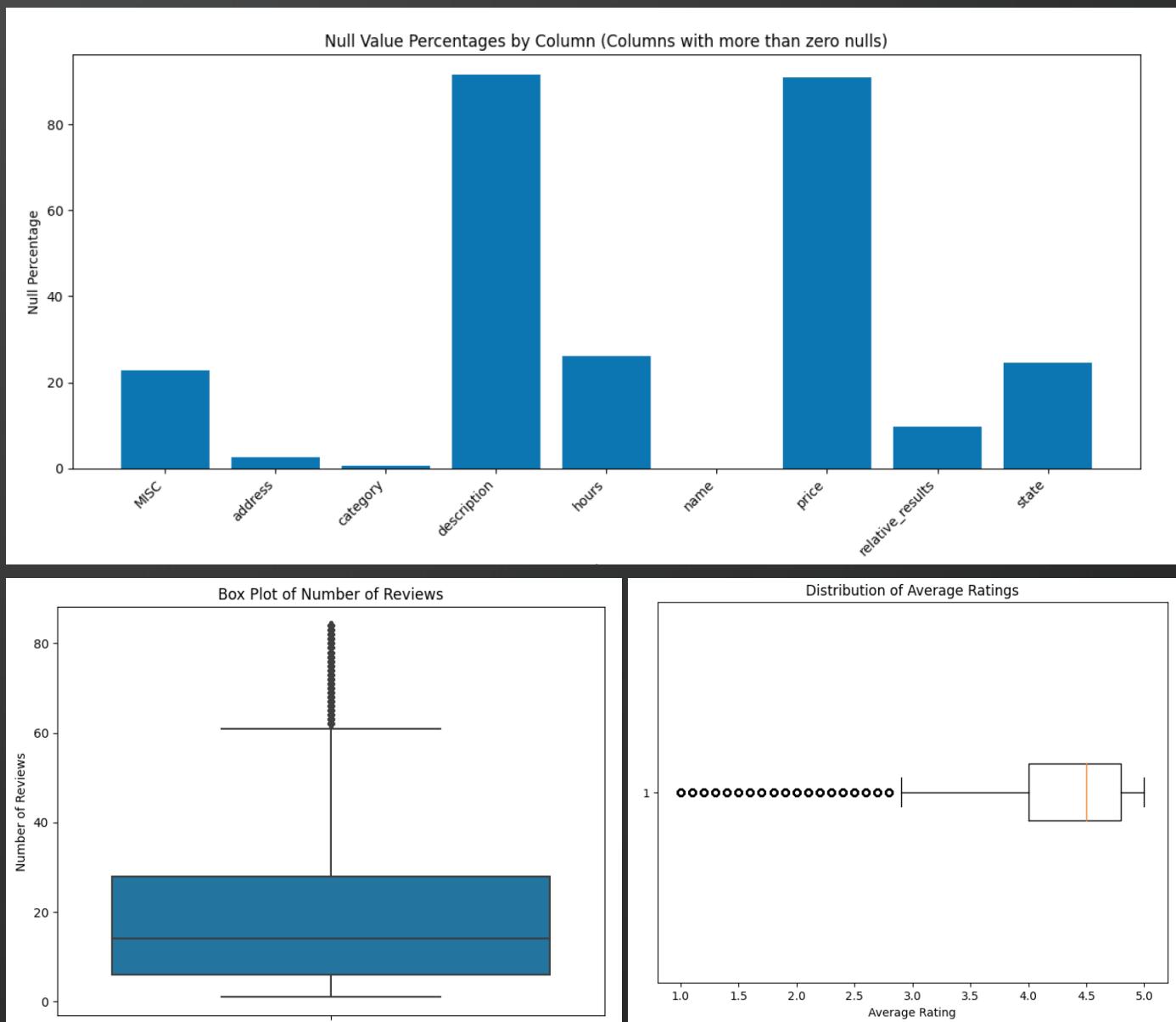
tip.json

Tips (consejos) escritos por el usuario. Los tips son más cortas que las reseñas y tienden a dar sugerencias rápidas.

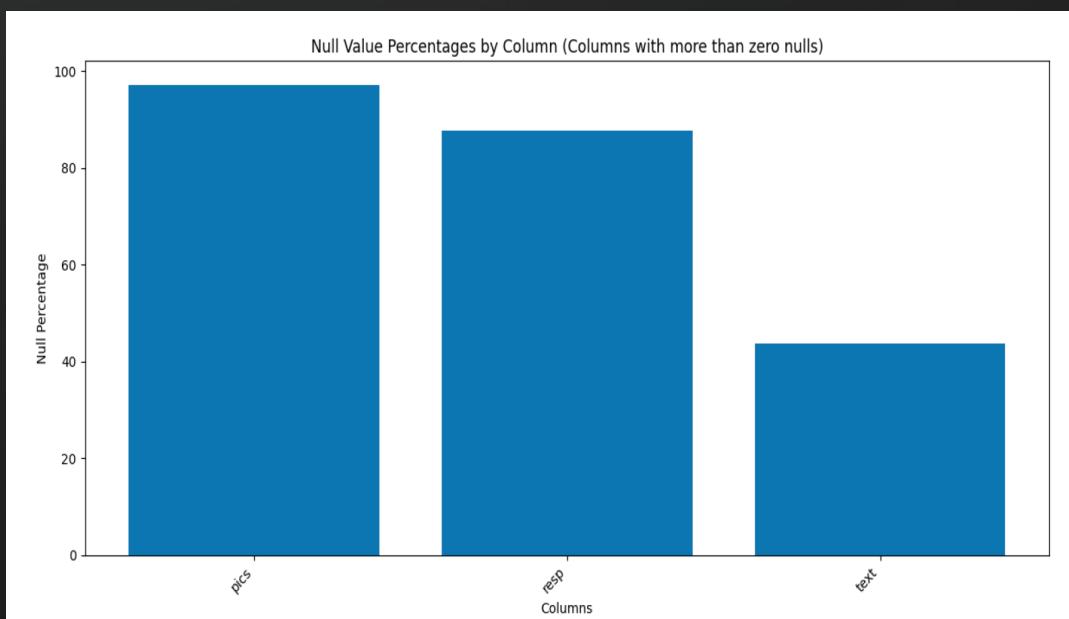
Columnas:

"text": string, texto del tip
"date": string, fecha cuando se escribió el tip YYYY-MM-DD
"compliment_count": entero, cuantos cumplidos totales tiene
"business_id": string, 22 caracteres, id del negocio que se refiere al negocio en business.json
"user_id": string, 22 caracteres de id de usuario, que se refieren al usuario en user.json

EDA Google Maps dataset metadata-sitios



reviews-estados



Profile report – 1

gmap1 – Job ID: 19693894

All Data

13 columns 3.00M rows 6 data types

● 93% valid values ● 0% mismatching values ● 7% missing values

hours

| Type | Array (Unknown) |
|---------------|-----------------|
| Valid | 2.22M |
| Mismatched | 0 |
| Empty | 779,513 |
| Top 20 values | |
| Closed | 2.48M |
| Monday | 2.22M |
| Tuesday | 2.22M |
| Friday | 2.22M |
| Sunday | 2.22M |
| Saturday | 2.22M |
| Wednesday | 2.22M |
| Thursday | 2.22M |
| Open 24 hours | 1.15M |
| 8AM-5PM | 1.08M |
| 9AM-5PM | 924,540 |
| 9AM-6PM | 468,607 |
| 10AM-6PM | 397,997 |
| 8AM-6PM | 323,816 |
| 10AM-7PM | 302,477 |
| 9AM-7PM | 259,394 |
| 10AM-8PM | 249,279 |
| 10AM-5PM | 246,858 |
| 8:30AM-5PM | 203,824 |
| 9AM-9PM | 189,497 |

MISC

| Type | Object (Schema)"Accessibility..." |
|-------------------|-----------------------------------|
| Valid | 2.32M |
| Mismatched | 0 |
| Empty | 683,048 |
| Top 18 values | |
| Accessibility | 1.93M |
| Service options | 899,744 |
| Planning | 781,534 |
| Amenities | 540,754 |
| Offerings | 415,458 |
| Payments | 397,861 |
| Health & safety | 380,731 |
| Highlights | 215,880 |
| Atmosphere | 190,657 |
| Crowd | 156,251 |
| Popular for | 151,535 |
| Dining options | 147,665 |
| From the business | 96,679 |
| Health and safety | 13,570 |
| Recycling | 1,354 |
| Getting here | 1,244 |
| Activities | 475 |
| Lodging options | 5 |

state

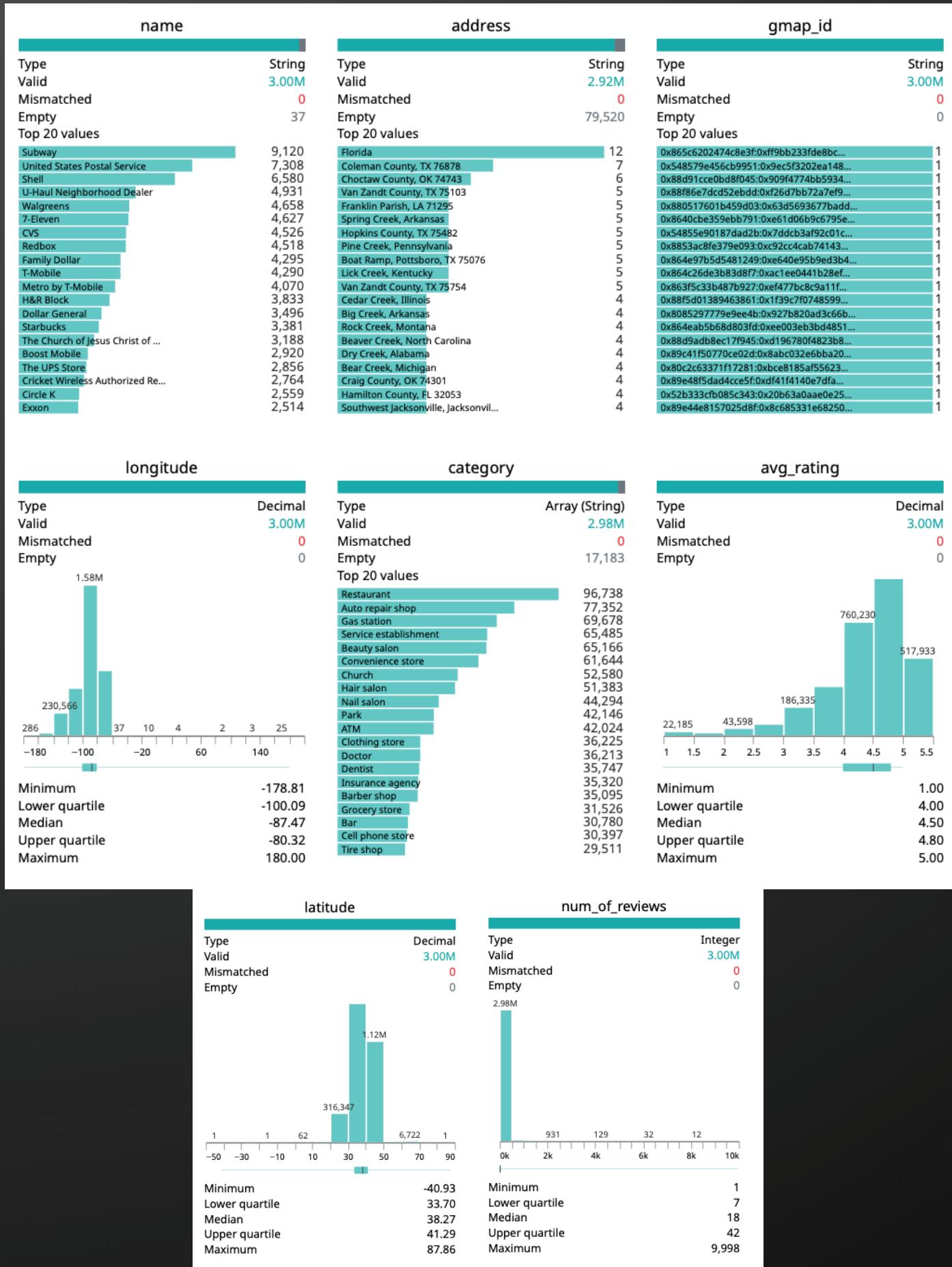
| Type | String |
|---------------------------|---------|
| Valid | 2.26M |
| Mismatched | 0 |
| Empty | 738,858 |
| Top 20 values | |
| Permanently closed | 189,166 |
| Open 24 hours | 154,761 |
| Closed : Opens 9AM | 135,817 |
| Closed : Opens 8AM | 121,668 |
| Open · Closes 5PM | 114,285 |
| Closed : Opens 8AM Mon | 102,561 |
| Closed : Opens 10AM | 94,559 |
| Closed : Opens 9AM Mon | 81,595 |
| Open · Closes 6PM | 67,905 |
| Open · Closes 7PM | 52,026 |
| Closed : Opens 11AM | 48,141 |
| Open · Closes 8PM | 46,237 |
| Open · Closes 9PM | 45,611 |
| Open now | 40,655 |
| Open · Closes 10PM | 39,912 |
| Closed : Opens 7AM | 36,665 |
| Closed : Opens 9AM Tue | 29,485 |
| Closed : Opens 8:30AM | 26,720 |
| Closed : Opens 10AM Mon | 25,511 |
| Closed : Opens 8:30AM Mon | 25,448 |

url

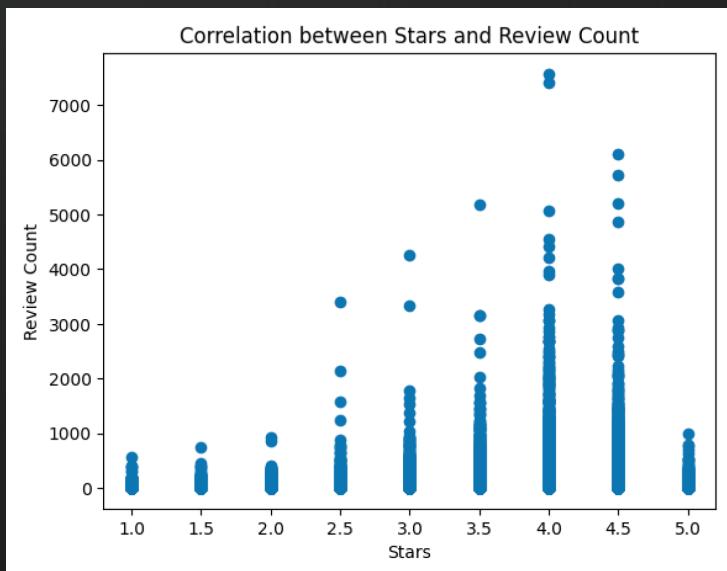
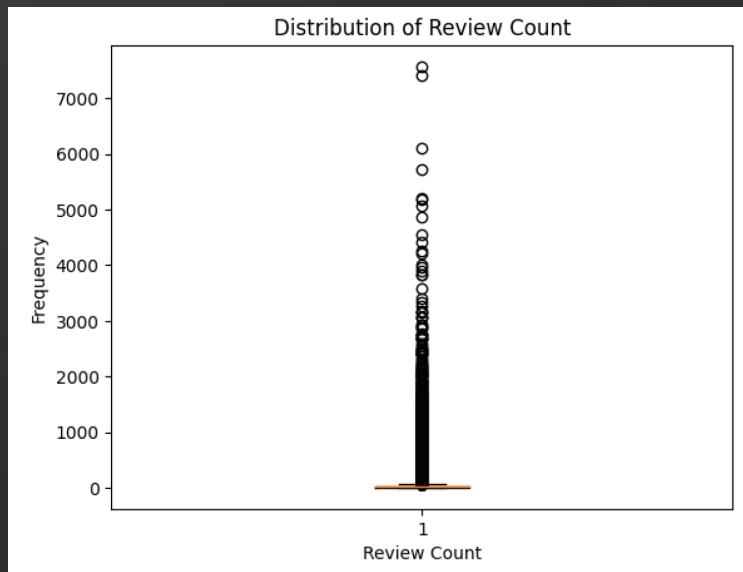
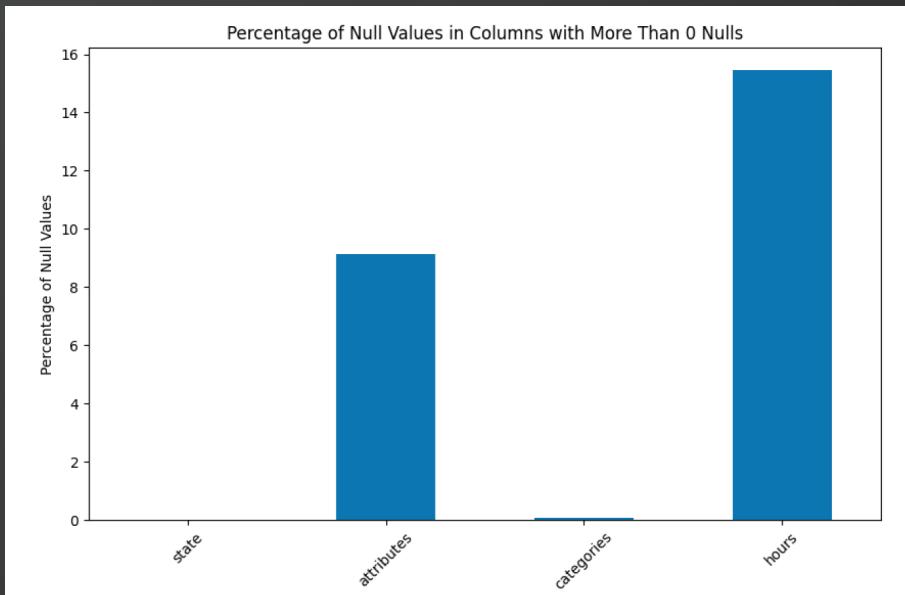
| Type | URL |
|---|-------|
| Valid | 3.00M |
| Mismatched | 0 |
| Empty | 0 |
| Top 20 values | |
| https://www.google.com/maps/place//... | 1 |

relative_results

| Type | Array (String) |
|--------------------------------------|----------------|
| Valid | 2.71M |
| Mismatched | 0 |
| Empty | 291,456 |
| Top 20 values | |
| 0x80e318c9493900b:0x56bc951dae5... | 118 |
| 0x874d7fba786afde1:0x6fcfae6cf57a... | 100 |
| 0x875302ca0fd80d5:0x771ea696eddd... | 82 |
| 0x880e2d40f11c2c41:0x45aebcc540e... | 79 |
| 0x8b2d03fe15ce15f:0x836ba716888... | 79 |
| 0x89c2573188a20c6f:0xb94193c5cf7... | 75 |
| 0x880fd211d77156cf:0xfc2c64b3b46... | 74 |
| 0x865cf60acd0c39f:0x3d90f2d6bf1... | 72 |
| 0x880e2d09c4d5d455:0x2f43b4b5dc1... | 68 |
| 0x880e32b36a64c051:0xb40b1e7f740... | 67 |
| 0x8640c5064c63dd4d:0x75d67c192c... | 67 |
| 0x88f50f6cab382ad:0x3c4a8b4addb... | 66 |
| 0x880e2d68e3bc1057:0xa2b27ded1e6... | 65 |
| 0x872b1472d6da105b:0x63070b09afc... | 64 |
| 0x874dbbb6d306f667:0xb7aa866af9f... | 64 |
| 0x89c259a210ab3365:0xe7a0d2fc40... | 61 |
| 0x89c2f3f878ebc8c7:0x3a35d73d660f... | 61 |
| 0x88f501784d281449:0xc15e34fe43c... | 60 |
| 0x89c258585a33dfb7:0xd9262685a9... | 58 |
| 0x89c25fc90de04785:0x2ecdbe825a4... | 57 |



EDA Yelp! dataset business.pkl



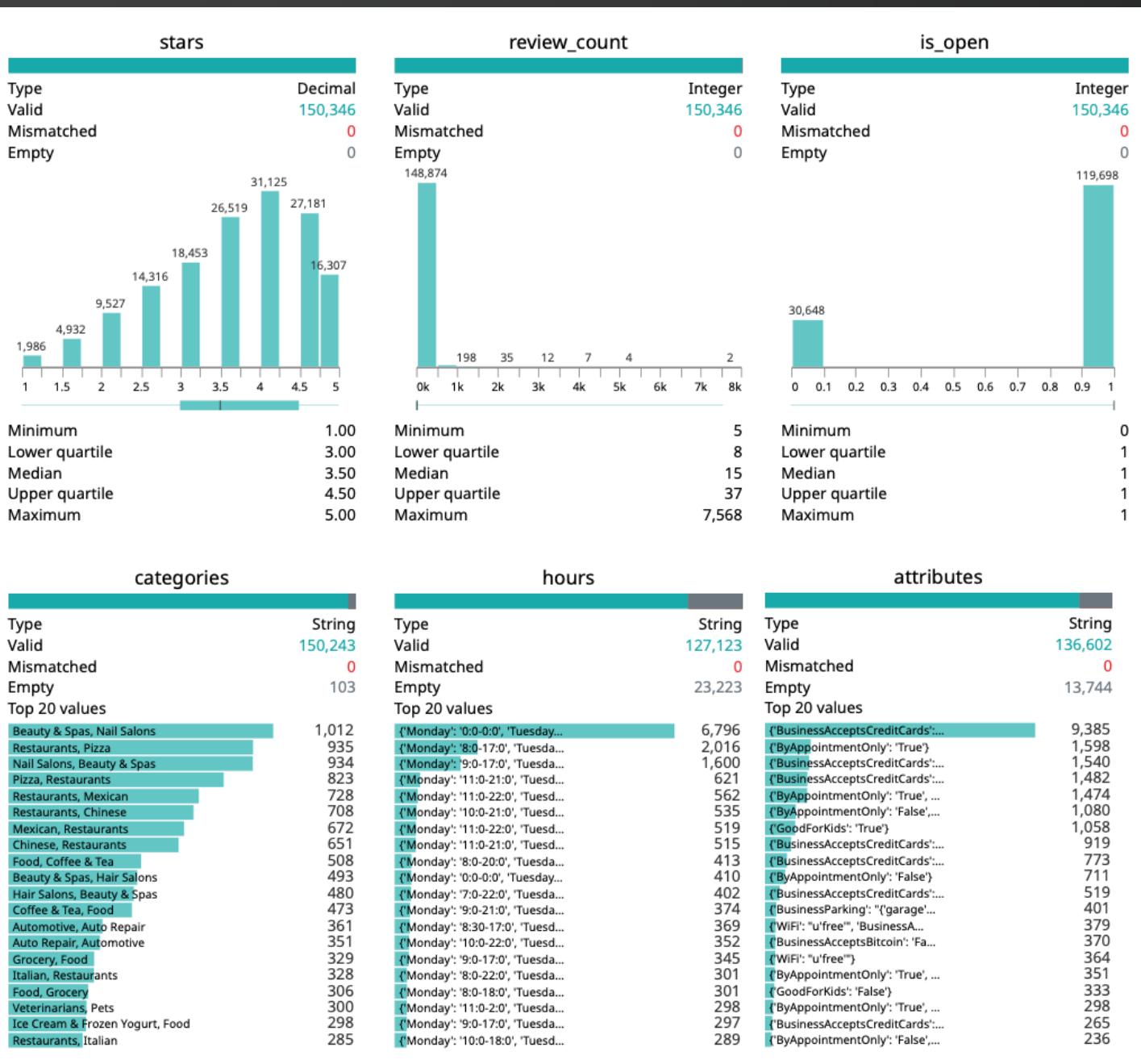
Profile report – business

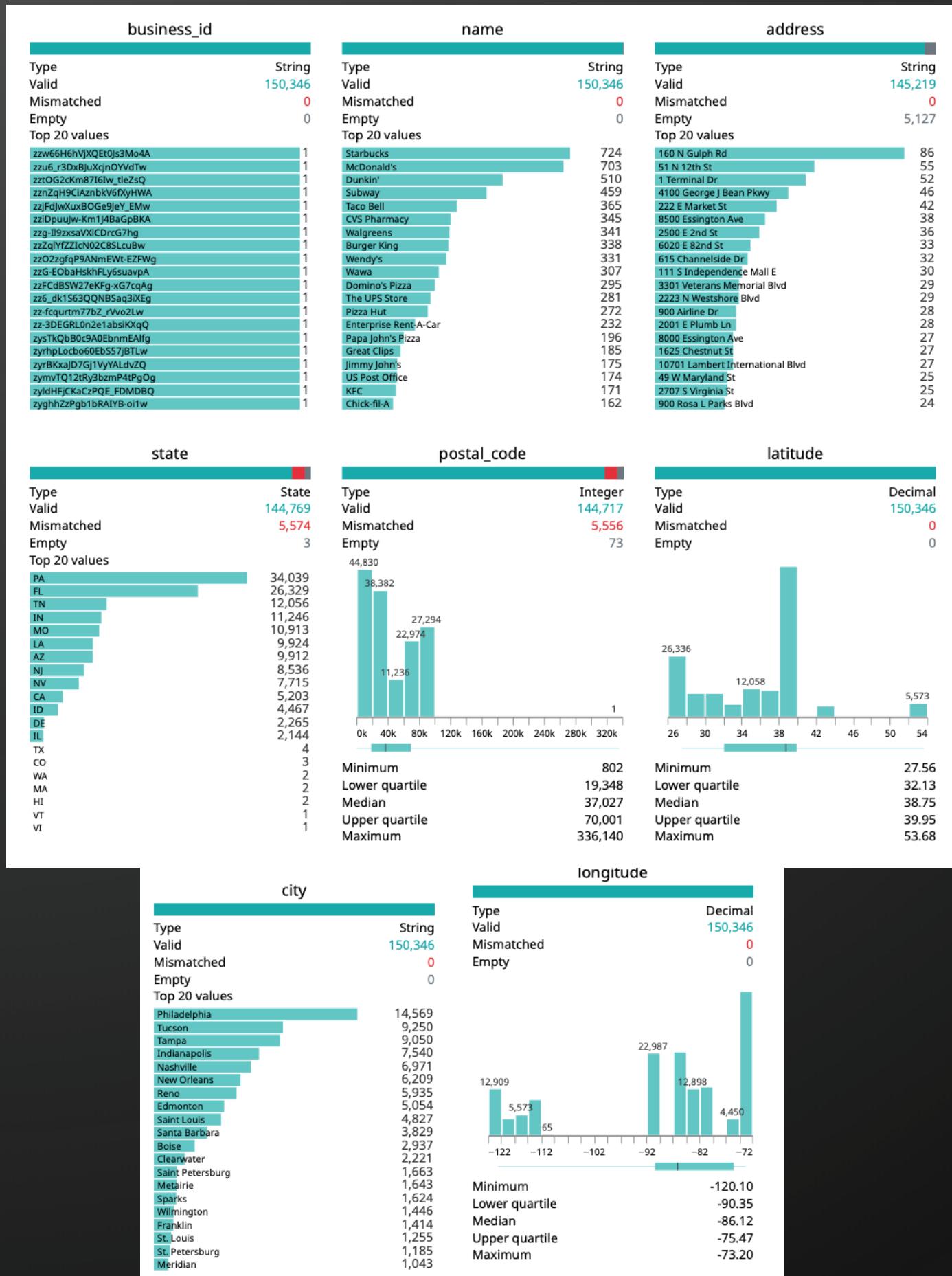
From Cloud Storage to BigQuery – Job ID: 19694647

All Data

14 columns 150,346 rows 4 data types

● 97% valid values ● 1% mismatching values ● 2% missing values





Durante esta semana de análisis de datos, hemos obtenido una perspectiva valiosa sobre la intersección entre los datos y el negocio. Hemos descubierto que los datos son una fuente inagotable de información y oportunidades para impulsar el crecimiento y la toma de decisiones inteligentes.

En primer lugar, hemos comprendido la importancia de tener datos limpios y ordenados. Al procesar y analizar datos de calidad, podemos obtener una visión precisa y confiable de nuestro negocio. Esto nos permite identificar patrones, tendencias y áreas de mejora que pueden tener un impacto significativo en nuestros resultados.

Además, hemos aprendido que el análisis de datos nos brinda una ventaja competitiva. Al utilizar herramientas y técnicas avanzadas, podemos extraer información profunda y reveladora de nuestros datos. Esto nos permite descubrir insights ocultos, identificar oportunidades de crecimiento y optimizar nuestros procesos para lograr mejores resultados.

GRACIAS 

