SPRINT #2

Introducción

En el siguiente documento se establece la estructura y el modelo de datos que utilizará nuestra aplicación final. Hemos optado por utilizar la plataforma de Google Cloud para el almacenamiento y la gestión de nuestros datos, eligiendo Google BigQuery como nuestro sistema de almacenamiento en la nube. Esta elección se basa en varias ventajas clave que ofrece BigQuery, como su capacidad de escalabilidad, rendimiento y facilidad de uso. Además, BigQuery es un servicio totalmente administrado, lo que significa que nos libera de la tarea de gestionar la infraestructura subyacente, permitiéndonos enfocarnos en el análisis de datos y la toma de decisiones.

- En primer lugar, se describe la plataforma tecnológica seleccionada, así como el flujo de datos desde su origen hasta su preparación en la solución propuesta.
- En segundo lugar, se detallan los procesos de automatización que respaldan el ciclo de vida de los datos.
- Por último, se proporciona información sobre la estructura de los datos y las relaciones entre los diferentes conjuntos de datos.

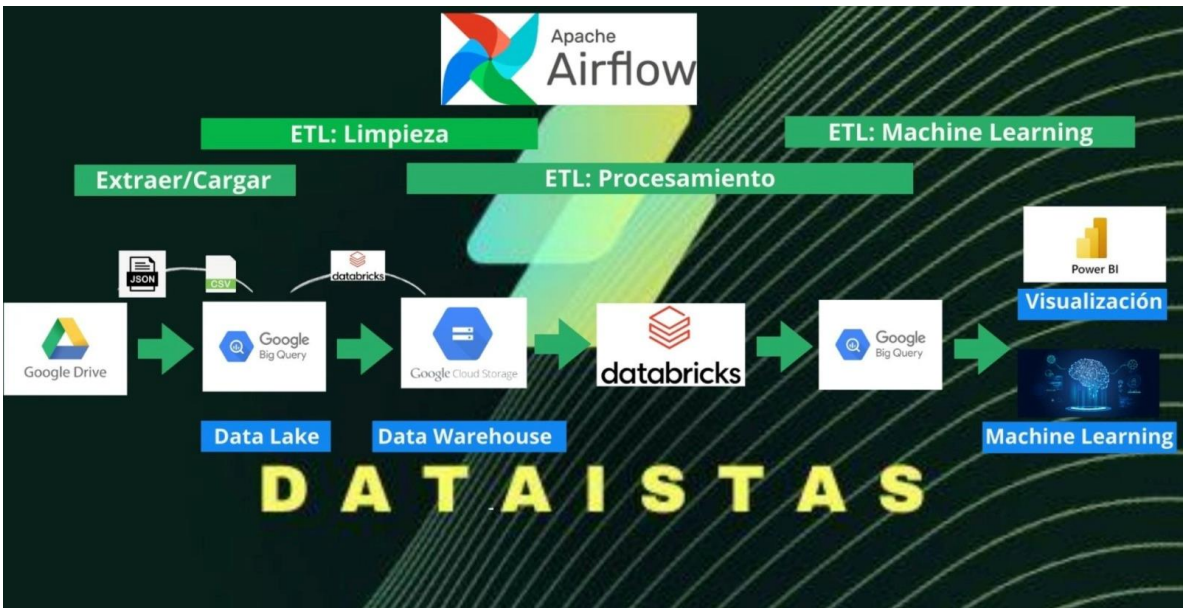
Tecnologías: Google Cloud Platform

Google Cloud Platform es una plataforma de servicios en la nube ofrecida por Google, la cual, proporciona una extensa variedad de herramientas y servicios que incluyen: almacenamiento, procesamiento, análisis, inteligencia artificial, aprendizaje automático, bases de datos y más. Estos servicios han sido diseñados para ofrecer una infraestructura confiable y escalable destinada a la ejecución de aplicaciones y el almacenamiento de datos en la nube.

En nuestro proyecto, hemos optado por utilizar Google Cloud Platform por diversas razones. En primer lugar, esta plataforma nos brinda una infraestructura altamente confiable, respaldada por las considerables inversiones de Google en su infraestructura subyacente, lo que garantiza una alta disponibilidad y una sólida protección de datos. Esto significa que podemos confiar en la disponibilidad y seguridad de nuestros datos y servicios en todo momento. Además, Google Cloud Platform ofrece una escalabilidad excepcional, permitiéndonos aumentar o reducir nuestros recursos de manera sencilla según nuestras necesidades. Esto resulta especialmente relevante en nuestro proyecto, donde trabajaremos con grandes volúmenes de datos y llevar a cabo operaciones de procesamiento intensivo.

Otra ventaja importante de Google Cloud Platform radica en su amplio abanico de servicios y herramientas. Contamos con acceso a servicios como BigQuery, Cloud Storage, Dataflow, Composer, entre otros, que nos permiten llevar a cabo tareas como almacenar, procesar, analizar y orquestar datos de manera eficiente. Esto facilita la construcción de soluciones integrales y escalables sin la preocupación por la infraestructura subyacente.

Flujo de datos



A continuación, presentamos el orden de pasos para la puesta en marcha del flujo de datos:

1. Se inicia configurando la aplicación de Drive en Google Cloud, creando un servicio con los permisos necesarios. Tomar en cuenta que se necesita la autentificación utilizando las credenciales adquiridas para acceder a la aplicación de Drive.
2. Identificamos los archivos específicos en Google Drive que deseamos extraer los archivos correspondientes para Google Maps y Yelp.
    a. **Google Maps:** Metadata y archivos de los estados correspondientes Georgia, California, Colorado, Nueva York y Texas.
    b. **Yelp:** Bussines, Review, Tip, User
3. Creación del datalake utilizando Big Query.
4. Conexión con Databricks para el procesamiento de los datos (limpieza). Los datos se someten a un proceso de limpieza antes de ingresar al almacén, lo que garantiza la consistencia y la calidad de los datos.
5. Creación del data warehouse en Google Cloud Storage.

EDA

Las actividades realizadas proporcionan una comprensión profunda de la información contenida en el conjunto de datos tanto para Yelp como para Google Maps, lo que facilita la toma de decisiones informadas y la identificación de patrones valiosos para el proyecto en cuestión.

Este análisis exploratorio se realizó de manera local antes de la limpieza de los datos. La información se encuentra en la carpeta titulada “1_EDA”.

1. **Tipo de dato para cada columna:** En esta etapa, se examinan todas las columnas del conjunto de datos para determinar qué tipo de datos contienen.

2. **Número de nulos por columna:** Se realiza un conteo de la cantidad de valores faltantes o nulos en cada columna.

3. **Exploración de variables categóricas.**

4.**Análisis de relaciones**: En esta etapa, se investigan las relaciones entre las diferentes variables del conjunto de datos. Esto incluye un análisis de correlación entre variables numéricas.

5. **Calificaciones por restaurante:** Evaluación del desempeño de cada restaurante. mediante la agrupación de datos por restaurante, cálculo de estadísticas como promedios de calificaciones y la creación de visualizaciones para comparar el rendimiento de los restaurantes.

LIMPIEZA DE DATOS

A continuación, se enlistan las actividades o procedimientos realizados que constituyen la etapa de Limpieza y Transformación de Datos.

1. **Eliminación de Valores Nulos en la Columna de Review de Usuario:** En la primera etapa de limpieza de datos, se procedió a eliminar los valores nulos que se encontraban en la columna de "review de usuario". Esto se hizo para asegurar la integridad de los datos y evitar problemas futuros en el análisis
   .
2. **Eliminación de Columnas con Altos Porcentajes de Nulos:** Después de una revisión exhaustiva de los datos de Google Maps, se identificó que las columnas "pics" y "resp" contenían más del 80% de valores nulos. En consecuencia, se tomó la decisión de eliminar estas dos columnas, ya que su contribución al análisis era limitada debido a la falta de datos significativos.

3. **Conversión de la Columna 'time' en Formato de Fecha:** Para facilitar el análisis temporal de las revisiones de los usuarios, se realizó la conversión de la columna 'time' al formato de fecha. Posteriormente, se crearon tres nuevas columnas que contenían la información del año, mes y día en que se registró la revisión del usuario.

4. **Selección de Datos para el Período de 2017 a 2021:** Se procedió a seleccionar las review de usuarios que estuvieran registradas en el período comprendido entre los años 2017 y 2021. Esto permitió enfocar el análisis en un rango de tiempo específico.

5. **Verificación de Duplicados y Eliminación:** Se llevó a cabo una verificación exhaustiva de duplicados en los datos y se procedió a eliminarlos. Esto aseguró que cada revisión de usuario fuera única en el conjunto de datos.

Procesamiento de Texto en la Columna de Revisión de Usuario:

6. **Conversión del Texto a Minúsculas:** Para homogeneizar el texto y facilitar el análisis de texto, se realizó la conversión de reviews a minúsculas.
7. **Eliminación de Emojis Utilizando la Librería "Demoji":** Se implementó la librería "Demoji" para eliminar los emojis contenidos en el texto de las revisiones de usuario. Esto contribuyó a una limpieza adicional del contenido de texto.
8. **Verificación y Traducción a Inglés:** Se verificó que todas las revisiones de usuario estuvieran en inglés. Aquellas revisiones que no estaban en este idioma fueron traducidas al inglés para mantener la consistencia en el análisis de texto.
9. **Eliminación de Signos de Puntuación:** Por último, se procedió a eliminar los signos de puntuación contenidos en el texto de las revisiones de usuario. Esto simplificó el análisis de texto y facilitó la extracción de información relevante.

Estas etapas de limpieza y transformación de datos fueron esenciales para preparar los datos de manera adecuada y asegurar que estuvieran listos para el análisis posterior. La uniformidad de los datos fue mejorada significativamente a lo largo de este proceso.

CLASIFICACIÓN DEL TIPO DE ESTABLECIMIENTO.

Una vez finalizada la limpieza de los datos, se procede a realizar la clasificación según el tipo de establecimiento.

1. Se optó por no eliminar las filas que contenían nulos para la columna 'Category' por lo que existen elementos nulos. Se sustituyen los nulos por ""

2. **Tokenización de la Columna Review**: Se desarrolla una función para tokenizar el texto en la columna de revisión. La tokenización implica dividir el texto en unidades más pequeñas, como palabras individuales o frases, para un análisis más detallado.

3. **Aplicación del Modelo para Identificar Palabras Clave:** Se utiliza un modelo de análisis de texto para identificar las palabras que se repiten con mayor frecuencia en la columna de comentarios.

4. **Creación de un Diccionario de Clasificación:** Se establece un diccionario para clasificar los tipos de comentarios en diversas categorías como: ***Restaurante, Agencia, Fotógrafo, Auto, Belleza & Spa, Mascotas, Salud, Iglesia, Casa, Oficina, Escuela, Banco, Deportes, Atracción Turística, Tienda, Sin categoría, Hospedaje, Fábrica***. Cada comentario se etiqueta en función de la categoría que mejor se ajuste a su contenido.

Este proceso de clasificación según el tipo de establecimiento ayuda a catalogar el tipo de establecimiento. Más adelante será de utilidad para seleccionar los establecimientos relacionados al sector gastronómico que son con los que trabajaremos.

ANÁLISIS DE SENTIMIENTOS

El análisis de sentimientos, también conocido como análisis de opiniones o minería de opiniones, es una técnica de procesamiento de lenguaje natural (NLP) que se utiliza para determinar la actitud, emoción o tono expresado en un fragmento de texto, como una revisión de producto, un comentario en redes sociales, un artículo de noticias o cualquier otro tipo de texto escrito. El objetivo principal del análisis de sentimientos es identificar y clasificar la polaridad emocional del texto en tres categorías fundamentales:

1. Positivo: Cuando el texto expresa una opinión favorable, satisfacción o emociones positivas hacia un tema o entidad.
2. Negativo: Cuando el texto expresa una opinión desfavorable, crítica o emociones negativas hacia un tema o entidad.
3. Neutral: Cuando el texto no muestra una carga emocional significativa o no expresa una opinión claramente positiva o negativa.

El análisis de sentimientos utiliza algoritmos de procesamiento de lenguaje natural y aprendizaje automático para analizar el texto y determinar la polaridad emocional. Estos algoritmos pueden tener en cuenta una variedad de características lingüísticas, como palabras clave, contexto gramatical, uso de negaciones, etc. A continuación, se enlistan los procedimientos que constituyen el análisis de sentimientos:

Una vez limpio la columna que contiene las reviews de los usuarios, se utiliza la librería TextBlob para el análisis de sentimientos.

CATEGORIZACIÓN DEL ANÁLISIS DE SENTIMIENTO

Partiendo del análisis de sentimiento hecho en el punto anterior, se procede a realizar la categorización de los resultados obtenidos.

1. Se crea una función para categorizar el sentimiento encontrado. Siendo:
    a. Positivo mayor a cero
    b. Neutral igual a cero
    c. Negativo menor a cero

2. El resultado del análisis de sentimiento son 2 columnas: la primera contiene el número asociado al utilizar TextBlob y la segunda corresponde a la categorización del sentimiento mencionada en el punto anterior.

CLASIFICACIÓN DE COMENTARIOS

En este proceso de clasificación de comentarios, se llevan a cabo una serie de pasos esenciales para analizar y categorizar las opiniones y reseñas de los clientes permitiendo identificar los temas y tendencias más relevantes en las reseñas. El objetivo principal es organizar y comprender de manera eficiente el contenido de las reseñas para beneficio de la empresa.

1. **Eliminación de Stop Words utilizando NLTK:** Se realiza la eliminación de las palabras de parada (stop words) en la columna de revisión (review) utilizando la librería NLTK. Esto ayuda a reducir el ruido en el texto y a centrarse en las palabras clave significativas.

2. **Tokenización de la Columna Review**: Se desarrolla una función para tokenizar el texto en la columna de revisión. La tokenización implica dividir el texto en unidades más pequeñas, como palabras individuales o frases, para un análisis más detallado.

3. **Aplicación del Modelo para Identificar Palabras Clave:** Se utiliza un modelo de análisis de texto para identificar las palabras que se repiten con mayor frecuencia en la columna de comentarios.

4. **Creación de un Diccionario de Clasificación:** Se establece un diccionario para clasificar los tipos de comentarios en diversas categorías como: ***limpieza, comida, tiempo, precio, servicio, estacionamiento y mascotas***. Cada comentario se etiqueta en función de la categoría que mejor se ajuste a su contenido.

Este proceso de clasificación de comentarios ayuda a organizar y comprender mejor el contenido de las reseñas de los clientes, lo que puede ser valioso para tomar decisiones informadas y realizar mejoras específicas en el negocio.

KPI

En el contexto de la mejora y crecimiento continuo de un negocio, es esencial establecer indicadores clave de rendimiento (KPI, por sus siglas en inglés) que permitan medir y evaluar el progreso hacia los objetivos establecidos. En este documento, se presentan tres KPI fundamentales que desempeñarán un papel crucial en la estrategia de mejora y expansión de la empresa. Estos KPI abordan aspectos clave, como la limpieza de los locales, la visibilidad en línea y la percepción del cliente.

- **KPI #1: Lograr una mejora del 10% en las reseñas relacionadas con la limpieza de los locales en los próximos 6 meses**, se deben implementar medidas específicas de limpieza y mantenimiento en los establecimientos. La supervisión constante y la retroalimentación de los clientes serán clave para lograr esta mejora del 10% en las reseñas de limpieza.

- **KPI #2: Incrementar en un 5% mensual el número de reseñas en la plataforma de Yelp Y Google Maps en los próximos 9 meses.** no solo mejora la visibilidad y la reputación en línea de un negocio, sino que también puede influir en las decisiones de compra de los consumidores y generar un ciclo virtuoso de retroalimentación positiva y crecimiento empresarial. Por lo tanto, puede ser una estrategia valiosa para mejorar la presencia en línea y el éxito a largo plazo de un negocio.

- **KPI #3: Crecer el número de reseñas positivas en un 6% en los locales en los próximos 5 meses.** No sólo mejora la percepción del negocio, sino que también tiene el potencial de aumentar las ventas, la confianza del cliente y la lealtad a la marca. Estas reseñas positivas pueden actuar como una poderosa herramienta de marketing y contribuir al éxito a largo plazo del negocio.

ANÁLISIS DE LA INFORMACIÓN OBTENIDA

Este análisis fue realizado une vez hecho la limpieza y el ordenamiento de los datasets para los restaurantes elegidos: Mc Donald's, KFC, Subway, Domino's Pizza

A continuación, se presentan las características que se tomaron en cuenta para dicha exploración:

1. **Total de Reviews por restaurante:** Este punto implica calcular la cantidad total de reseñas u opiniones que ha recibido cada restaurante en el conjunto de datos. Puede ser útil para identificar los restaurantes más y menos comentados.

	Cadena	Total_Review_Count
3	Starbucks	583563
2	McDonald's	570342
0	Domino's Pizza	161012
1	KFC	65028
4	Subway	33431

2. **Cantidad de Restaurantes por cadena:** En este caso, se realiza un conteo de la cantidad de restaurantes que pertenecen a cada cadena o franquicia presente en el conjunto de datos.

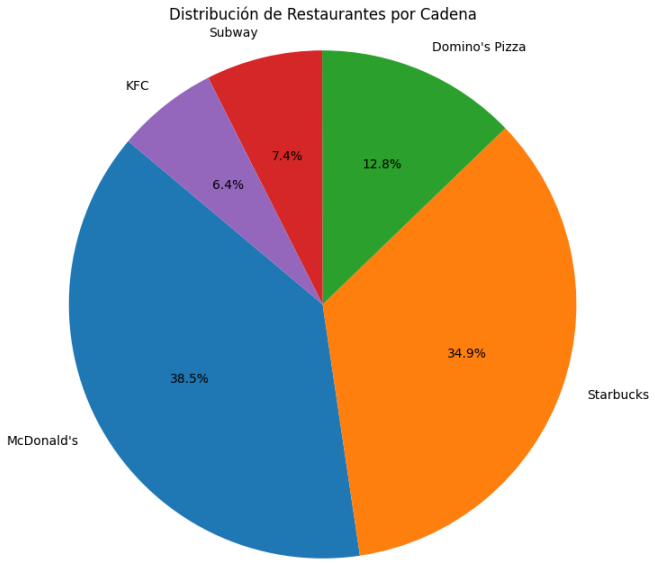
Esto proporciona una visión de la distribución de restaurantes entre diferentes cadenas y puede ayudar a identificar cuáles tienen una mayor presencia.

	Cadena	Cantidad de Restaurantes
0	McDonald's	14653
1	Starbucks	13284
2	Domino's Pizza	4884
3	Subway	2828
4	KFC	2454

3. **Cantidad de restaurantes por estado:** Se realiza un conteo del número de restaurantes por cadena para cada estado.

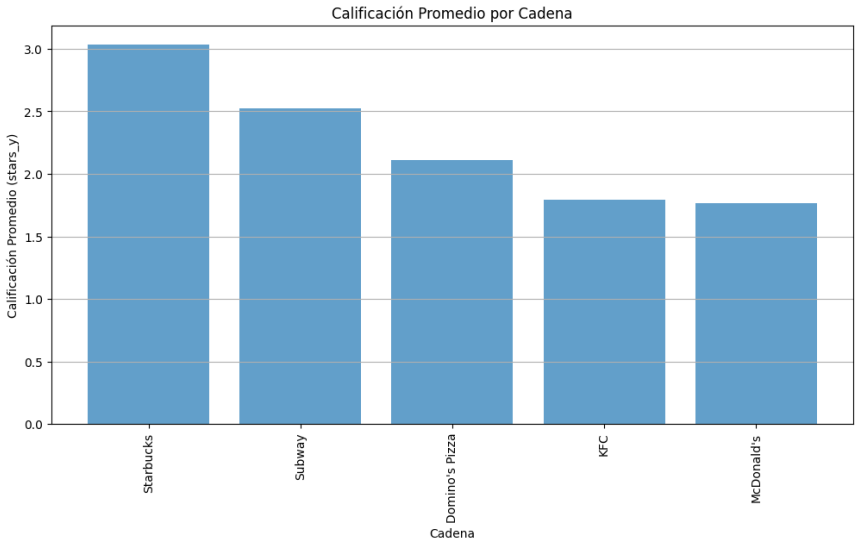
name	Domino's Pizza	KFC	McDonald's	Starbucks	Subway
state					
AB	262	75	298	404	89
AZ	270	88	956	814	191
CA	121	60	512	545	144
DE	32	3	141	329	36
FL	770	395	2129	2069	621
ID	177	57	419	307	87
IL	50	12	217	242	72
IN	369	200	1137	1069	255
LA	532	244	804	1070	136
MO	352	108	1259	967	135
NJ	332	171	1049	822	161
NV	224	184	604	753	92
PA	982	574	3762	3003	579
TN	411	283	1366	890	230

4. **Porcentaje de restaurantes para cada cadena:** Se calcula el porcentaje que representa cada cadena o franquicia en relación con el total de restaurantes en el conjunto de datos. Ayuda a entender la proporción de restaurantes que pertenecen a cada cadena en el contexto general y es útil para comprender la participación de mercado de cada una.



5. **Promedio de Calificaciones por Restaurante y Estado:** Esto proporciona información sobre la satisfacción promedio de los clientes en cada estado para cada restaurante individual. Es útil para identificar cómo varía la calidad de los restaurantes según su ubicación.

6. **Promedio de Calificaciones por Restaurante:** Gráfico sobre la satisfacción promedio de los clientes para cada restaurante individual. Es útil para identificar cómo varía la calidad de los restaurantes.

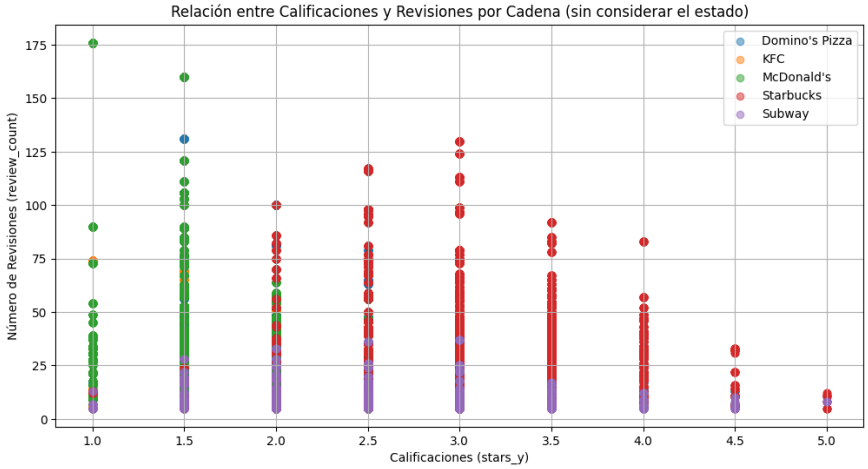


7. **Distribución de Calificaciones por Estado y Cadena de Restaurantes:** Se examina la distribución de las calificaciones otorgadas por los clientes en relación con el estado en el que se encuentra el restaurante y la cadena a la que pertenece. Esto permite comprender

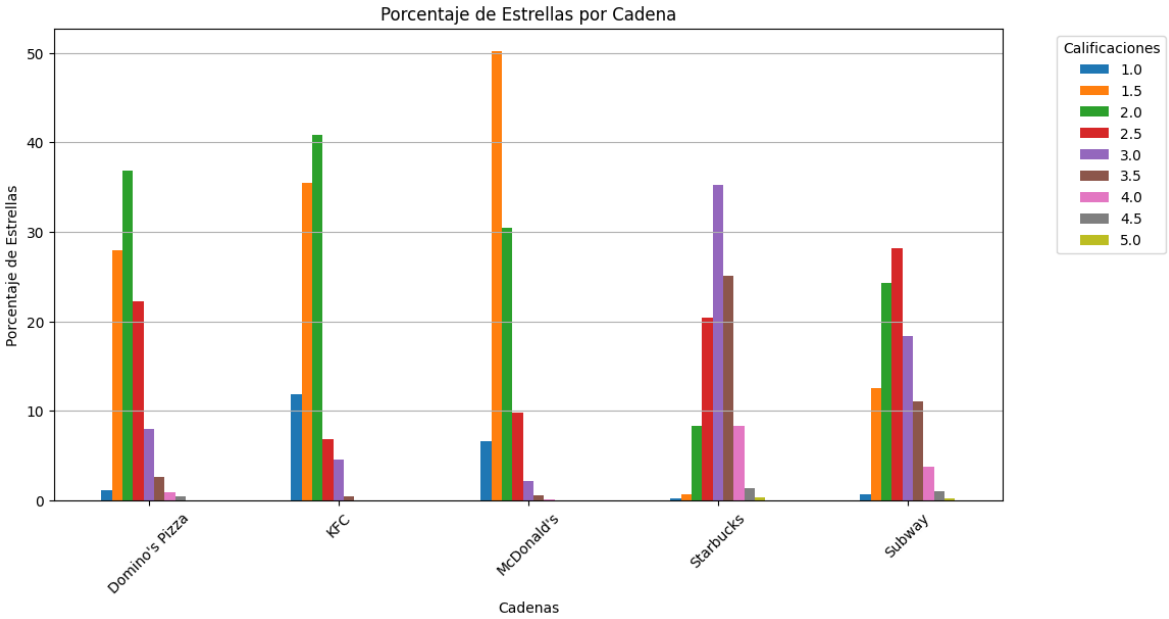
cómo se distribuyen las calificaciones tanto a nivel geográfico como dentro de las diferentes cadenas. Puede ayudar a ***identificar tendencias regionales*** y evaluar el desempeño de cada cadena en diferentes ubicaciones.

8. **Relación entre calificaciones y Revisiones por Cadena (sin considerar estado):** Este análisis busca examinar la relación entre las calificaciones de los restaurantes y la cantidad de revisiones o reseñas que han recibido, sin tener en cuenta la ubicación geográfica (estado). Esto para ayudar a comprender si la popularidad o la calidad de una cadena influye en la cantidad de reseñas que recibe.

stars_y	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
name									
Domino's Pizza	55	1365	1800	1085	387	125	44	23	0
KFC	290	872	1002	169	111	10	0	0	0
McDonald's	972	7359	4459	1438	323	84	18	0	0
Starbucks	31	95	1102	2716	4688	3337	1100	179	36
Subway	20	355	686	796	518	312	106	29	6



9. **Porcentaje de Estrellas por cadena**: En este punto, se calcula el porcentaje de estrellas otorgadas a cada cadena de restaurantes en relación con el total de estrellas otorgadas en el conjunto de datos. Esto proporciona una visión de cómo se distribuyen las calificaciones por estrellas entre las diferentes cadenas, lo que puede ayudar a identificar cuáles cadenas tienen una mayor proporción de calificaciones positivas o negativas en comparación con otras.


Porcentaje de Estrellas por Cadena
Porcentaje de Estrellas
0
10
20
30
40
50
Domino's Pizza
KFC
McDonald's
Starbucks
Subway
Cadenas
Calificaciones
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0