Documentación

### Introducción

Este proyecto tiene como objetivo desarrollar un análisis integral de la industria de la comida americana en Florida. El propósito principal es obtener información útil para la definición de estrategias de negocios, mediante la creación de KPIs y modelos de Machine Learning (ML) que reflejen con precisión la situación del mercado. El proceso de trabajo sigue las fases tradicionales de un análisis de datos: ETL (Extracción, Transformación y Carga), EDA (Análisis Exploratorio de Datos), automatización en AWS, modelado de ML y la creación de un dashboard interactivo en Power BI. A continuación, se detallan los pasos clave realizados en el análisis de datos.

### ETL-EDA de reviews\_Google

El análisis comenzó con el procesamiento de más de 2.250.000 reseñas de Google Maps del estado de Florida, originalmente en formato JSON. Los datos fueron transformados a CSV para facilitar su manejo. Se exploró la calidad del dataset, identificando duplicados y valores nulos, y se eliminaron columnas irrelevantes con más de un 80% de datos faltantes. Posteriormente, se realizó un análisis exploratorio que incluyó la distribución de las calificaciones, la detección de duplicados y un análisis de términos frecuentes en las reseñas mediante una nube de palabras.

### ETL-EDA de datos de “check-ins”

En cuanto al archivo de check-ins, que contenía el ID del negocio y la hora de cada check-in, se transformó el JSON original a un CSV y se realizó una exploración de datos similar. Se normalizaron fechas, se crearon índices secuenciales, y se analizó la actividad de check-ins por año, hora del día, día de la semana y mes. Los resultados mostraron patrones de mayor actividad durante los fines de semana, especialmente en las tardes y noches.

### ETL-EDA del archivo “business” de Yelp

El análisis del archivo "business" de Yelp incluyó la transformación de datos de formato "pickle" a CSV y la limpieza de columnas irrelevantes. Se desanidaron diccionarios dentro de la columna "attributes" para un análisis más detallado de los atributos de los negocios, y se examinó la columna "categories" para enfocar el análisis en restaurantes y tipos de comida. También se procesaron los horarios de apertura y cierre de los negocios, creando columnas específicas para cada día de la semana.

### ETL-EDA del archivo de “usuarios”

El archivo de usuarios, en formato parquet, fue transformado para garantizar la integridad de los datos, convirtiendo tipos de datos y eliminando duplicados. Se realizó un análisis de las columnas relacionadas con cumplidos, y se observó que la mayoría de los usuarios no reciben cumplidos con frecuencia. Finalmente, se creó una matriz de correlación para analizar relaciones entre variables numéricas y se exportó un archivo CSV comprimido para un manejo más eficiente de los datos.

### ETL-EDA del archivo de “tips”

El archivo de tips en formato JSON fue procesado de manera similar a los anteriores. Se transformaron los tipos de datos, se eliminaron duplicados y se separaron fechas y horas para facilitar el análisis. Al igual que en el análisis de usuarios, se eliminó la columna "compliment\_count" debido a su baja variabilidad. El archivo final se exportó en un formato comprimido para optimizar el uso de espacio.