# Data Pipeline Implementation Using Azure Services

## About This Project

This project showcases the design and implementation of a comprehensive data pipeline that processes raw data from Kaggle through an end-to-end solution built on Azure cloud services. The goal was to create a robust, scalable pipeline that automates data ingestion, transformation, storage, and analysis, ultimately enabling data-driven decision-making.

The pipeline begins with Azure Data Factory, which orchestrates the extraction of raw data from Kaggle. Once ingested, the data is stored in Azure Data Lake Storage. Azure Databricks is then employed to develop and execute Spark-based transformations, refining the raw data into a more usable form. The transformed data is then stored back into Azure Data Lake Storage for further processing. Finally, Azure Synapse Analytics is used to run SQL queries on the transformed data, uncovering insights and creating visualizations that drive informed decisions.

This project not only demonstrates the integration of multiple Azure services but also highlights the efficiency and scalability of cloud-based data pipelines.

## Key Features

- **Automated Data Ingestion**: Orchestrated data extraction from Kaggle using Azure Data Factory.
- **Scalable Data Storage**: Leveraged Azure Data Lake Storage for secure, scalable data storage.
- **Data Transformation**: Used Azure Databricks and Apache Spark to transform raw data into actionable insights.
- **Advanced Analytics**: Ran SQL queries and generated visualizations using Azure Synapse Analytics to support data-driven decision-making.

## Key Technologies

- **Azure Data Factory**: Orchestrates data movement and transformation, automating the data ingestion process from Kaggle.
- **Azure Data Lake Storage**: Provides scalable and secure data storage for both raw and transformed datasets.
- **Azure Databricks**: Utilized for data processing and transformation using Apache Spark, enabling fast and scalable data processing.
- **Apache Spark**: Core technology used within Azure Databricks for large-scale data transformation and analytics.
- **Azure Synapse Analytics**: Serves as the data warehouse and analytics engine for running SQL queries, analyzing data, and creating visualizations.
- **SQL**: Used to query and analyze the transformed data within Azure Synapse Analytics.