Outline the problem:

We aim to help restaurant businesses improve by using a sentiment analysis model that is used to analyze the Yelp reviews dataset (10k+ rows) which helps us understand what makes a positive and negative customer experience.

Outline your approach:

Initially, we uploaded the Yelp reviews dataset and removed a few columns (for example 'joy' which we assumed were reactions to reviews?). We then prepared the data using recipes. We used the sentiment analysis recipe. Then, we did topic modeling on a jupyter notebook to analyze the most popular words that relate to positive and negative reviews. We removed a lot of neutral words that would not add much information to the review (for example: "it", "and", "is", "food", "there", etc.). When we realized there were too many different results, we decided to split the restaurants by its cuisine. We also noticed there were a lot of words indicating a car dealership store which is when we realized the Yelp reviews dataset also included reviews for stores like car dealerships and hair and nail salons. We then found the Yelp business dataset to contain more information on the businesses such as their location and the type of businesses, then we filtered out the restaurants for our goal.

Using dataiku, we merged 2 datasets and prepared the datasets, split the dataset into types of restaurants (5 categories: café, Asian, nightlife, Italian, and seafood), and did sentiment analysis. Finally, we created a notebook for topic modeling and summarized our results with analysis.

Describe your dataset:

We obtained the dataset from Yelp's open-source website. We downloaded the Yelp reviews dataset (10k+ text reviews on 3,930 restaurants) and the Yelp business dataset (which includes information on the businesses such as location, type of business, etc.) and uploaded it onto dataiku.

Showcase impact:

The model would give us insight from 10k text reviews. It aids the improvement of 5 different types of restaurants (café, Asian, nightlife, Italian, and seafood) through behavior suggestions and dish recommendations. For example, the occasion (whether the restaurant is most popular for breakfast, lunch, or dinner, food types or flavors, environment, and preferred service). It highlights the positive and negative qualities of a large dataset of restaurants by selecting the top keywords that are most relevant from these reviews. This is a reusable pipeline that can be reused using other datasets.