DATA CURATION PROJECT: YELP DATASET ANALYSIS ***DATASET OVERVIEW***

DATASET NAME
Yelp Dataset (from Kaggle)

DOMAIN
The dataset represents user-generated reviews, business details, user attributes, and check-ins from the Yelp platform, providing a comprehensive insight into local businesses, primarily in the North American region.

RESOURCE
The dataset was procured from Kaggle's platform via the following link:

[Kaggle Yelp Dataset](https://www.kaggle.com/datasets/yelp-dataset/yelp-dataset/data).

CHARACTERISTICS
- Dataset Size: Comprising multiple CSV and JSON files.
- Attributes: Diverse attributes span across several files. For example, `yelp_business.csv` encompasses attributes like `address`, `categories`, `city`, `latitude`, `longitude`, `name`, `postal_code`, `review_count`, `stars`, among others.
- Data Format: Data is in both CSV and JSON formats, allowing for flexible data processing and analysis.

REASON FOR DATASET SELECTION

Comprehensive Data (Integration):
- The dataset has diverse details, from reviews to business information, making it good for combining or linking related data together. For example, the user's review can be connected with the specific business they reviewed using unique identifiers.

Relevance & Impact (Warehousing):
- The dataset's reviews can be organized neatly, helping in easily pulling out useful insights.

INITIAL DATA QUALITY ASSESSMENT

IDENTIFIED ISSUES
1. Missing Data: Potential absence of certain attributes, such as `postal_code` or `address` for businesses. Some reviews might be devoid of associated user attributes.
2. Inconsistencies: Potential variations in data notation, especially in textual fields. For instance, categorization of businesses might exhibit inconsistencies.
3. Noise: The presence of potentially non-informative data, such as businesses with scanty reviews or overly brief reviews.
4. Structural Challenges: The data, distributed across multiple files, might present integration challenges, necessitating careful merging based on unique identifiers like user or business IDs.

METHODOLOGY FOR ISSUE IDENTIFICATION

Descriptive Statistics: Using the VizierDB's interface, basic statistics on data columns can be computed to identify outliers and missing values.

Textual Analysis: The VizierDB's Python UDFs can be utilized to apply NLP techniques on reviews, helping spot inconsistencies in text.

Visual Inspection: Similar to NLP, Python UDFs in VizierDB can be used to generate simple visualizations, aiding in identifying data quality concerns.

Other Similar dataset - https://github.com/bookingcom/ml-dataset-mdt [Booking.com]