Overview

Context: You're planning to open a restaurant and have access to Yelp data. This project simulates a real data case study: acquiring semi-structured data, transforming it into a queryable database, and extracting business insights to inform a business decision.

Assessment: 20 points (Team project)

What You'll Deliver

One public Kaggle notebook (or Google Colab) per team that demonstrates:

1. **Data understanding** – Exploration of the Yelp JSON dataset structure, scope, and time period
2. **Database implementation** – All Yelp tables converted and loaded into your team's Azure SQL Server with appropriate metadata and constraints
3. **Business analysis** – (At least) 10 SQL queries that answer strategic questions about where and what type of restaurant to open

---

The Three Milestones

Milestone 1: Acquire & Explore the Data and setup up Azure DB (by Oct 27)

What to do:

- Create a Kaggle account and access the Yelp Dataset

⍰ Links to an external site.

⍰ Create a new notebook (each team member should practice this; you can add collaborators later)

⍰ Load and explore the JSON files – understand the structure, nested fields, data types, geographic coverage, and time period

⍰ Document your findings in the notebook with clear explanations

⍰ Setup your Azure database. Everyone can setup and try. Final submission will need only one members account to be used.

- Azure setup instructions : Demo Video Link

Link to Sign Up for free Azure Student Accounts

-
    - Links to an external site.
- At least two members from team must meet with the TA's during office hours before Oct 27 midnight to discuss progress for milestone 1 and 2. Check the https://canvas.illinois.edu/calendar and book a time to meet with either of the two TA's.

What we're looking for: Evidence that you understand the data before transforming it. Your notebook should show what you discovered about the dataset's structure and scope.

💡 **Common Challenge:** Memory errors when loading large files

Approach: Large datasets (like the 6M+ row Review table) can't be loaded all at once. Research Pandas' chunksize parameter for reading data in manageable pieces. The documentation and your exploration will guide you to the right approach.

Milestone 2: Build the Database

What to do:

- Connect to your team's Azure SQL Server
- Transform the JSON data into properly structured SQL tables
- Set appropriate data types, primary keys, and constraints
- Verify all tables loaded correctly with complete data

What we're looking for: A complete, well-structured database. Check your work against the reference yelp_champaign database (from Lab 1) to ensure proper metadata specifications.

💡 **Common Challenges:**

- **Nested JSON columns:** The Business table has 'attributes' and 'hours' columns that need special handling. You'll need to explicitly specify data types for these columns.
- **Multi Valued Attributes and Derived Attributes** : As discussed in class, these cannot be stored as is once you create your tables. Review Week 10 slides.

- **Primary keys:** Azure SQL requires explicit primary key definitions. Examine the Yelp data's ID columns – they're designed to be primary keys. Reference Yelp Data modelfor help
- **Memory management:** Use chunksize for uploads just like you did for reading data.

Milestone 3: Extract Business Insights

What to do:

Write (at least)10 SQL queries that help you make strategic decisions about opening a restaurant. Before writing queries, define your concept:

- Which city are you targeting and why?
- What type of restaurant? (cuisine, price point, style) and Why?
- Who is your target customer? and Why?

Your queries should address questions like:

- **Market analysis:** Where is the competition? How saturated is the market? What are competitive ratings?
- **Customer insights:** What do customers value? What are common complaints? What drives success?
- **Opportunity identification:** Which areas are underserved? What gaps exist? What would differentiate you?

What we're looking for: Queries that reveal non-obvious insights and inform real business decisions. Each query should have clear business context explaining what decision it informs.

⚠️ **Quality Check:** For each query, ask yourself:

- Does this answer inform a specific decision I need to make?
- Could I explain this insight to an investor?
- Does it reveal something I couldn't easily guess?

Submission Requirements

1. **Make your notebook public** before submitting

2. **One submission per team** with all collaborators listed
3. **After submission:** Review two other team submissions anonymously and provide constructive feedback within 24 hours of the deadline.
4. **Complete the peer review survey** to rate your contribution and your teammates' contributions within 24 hours of the deadline.