# **Yelp Business Data Analytics**

# Project Synopsis

# CDAC DBDA

**Guided By:**

**Ishan sir**

**Submitted By:**

**Group 2**

****

**Centre for Development of Advanced Computing (CDAC)** Post Graduate Diploma in Big Data Analytics (PG-DBDA) Mumbai Centre – USM VITA

**Academic Year: Feb – Aug 2025**

## Team Members:

The project team consists of seven dedicated and skilled members:

* Karan Anpat– Team Leader
* Mrunmai Jawalekar– Team Member
* Tushar Nandurkar – Team Member
* Mayuri Kakde – Team Member
* Yogeshwar Patil – Team Member
* Sanjana Dalvi – Team Member
* Santosh – Team Member

**Title:**

**Yelp Business Data Analytics**

Problem Statement:

This project focuses on exploring the Yelp dataset to uncover key patterns in customer reviews, business performance, and user engagement across various cities and categories. Through data exploration techniques, the goal is to identify trends in ratings, review activity, and customer behavior—without using any predictive modeling. The insights derived can help understand what drives user satisfaction and how businesses perform across different locations and types.

Project Objectives:-

- Uncover how star ratings vary across businesses and categories.

- Spot trends in review volume and customer sentiment over time.

- Highlight top businesses based on ratings and review count.

- Compare customer engagement across cities and business types.

- Analyze review behavior patterns like length, timing, and frequency.

**KPIs**

1. **Total Number of Businesses**
2. **Average Business Rating**
3. **Number of Business Categories**
4. **Number of Businesses with 5-Star Ratings**
5. **Number of Open vs Closed Businesses**
6. **Total Number of Users**
7. **Total Number of Reviews**
8. **Average Review Rating**
9. **Most Active Reviewers**

**Dataset Description:**

**Business Data**

* **What it contains**: Information about businesses such as **name**, **location** (address, city, state, postal code, latitude/longitude), **categories** (e.g., Restaurants, Beauty, Shopping), and **average star rating**.
* **Purpose**: Helps in identifying popular business categories, analyzing performance by location, and understanding business characteristics (e.g., hours, attributes).

**Review Data**

* **What it contains**: Full-length reviews written by users, including **review text**, **user ID**, **business ID**, **star rating**, and **timestamp**.
* **Purpose**: Essential for performing **sentiment analysis**, understanding customer opinions, extracting key feedback, and evaluating user satisfaction over time.

**User Data**

* **What it contains**: Profiles of users with attributes like **number of reviews written**, **average rating given**, **number of fans**, **friend connections**, and **elite status**.
* **Purpose**: Useful for identifying **highly active or influential users**, analyzing user engagement, and clustering behavior types.

**Tip Data**

* **What it contains**: Short user-submitted tips about businesses, often more concise than reviews. Includes **tip text**, **business ID**, **user ID**, **date**, and **compliment count**.
* **Purpose**: Ideal for quick opinion mining, keyword extraction, and analyzing what users commonly highlight or appreciate about a business.

**Check-in Data**

* **What it contains**: Timestamped check-in information that records **when users visited businesses**.
* **Purpose**: Valuable for identifying **peak hours**, **popular visiting times**, and **seasonal or daily traffic patterns**.

**Scope of Work:**

**1. Data Cleaning and Preprocessing**

* **Load and integrate datasets: Combine business, review, tip, checkin, and user datasets into a structured model.**
* **Handle missing or inconsistent data: Clean null values in fields like attributes, categories, and hours.**
* **Feature extraction: Extract fields such as city, state, category, review count, average stars, and check-in times for analysis.**

**2. Exploratory Data Analysis (EDA)**

* **User Behavior Analysis:**
  + **Track most active users based on review count, tip activity, and average rating.**
  + **Visualize user engagement levels and trends across cities or states.**
* **Business Performance Insights:**
  + **Identify top-rated businesses by category and location.**
  + **Analyze distribution of star ratings across business types (e.g., restaurants, salons, gyms).**
  + **Highlight businesses with high ratings but low review count (hidden gems).**
* **Review Trends:**
  + **Explore temporal trends in review submissions (monthly/yearly).**
  + **Detect peak review months or seasonal business performance.**
* **Geographic Trends:**
  + **Use Power BI maps to show:**
    - **Business density by location**
    - **Top-rated cities**
    - **Heatmaps of user activity**

**3. Tip and Check-in Pattern Visualization**

* **Compliment Insights:**
  + **Analyze tips with the highest compliment counts.**
  + **Identify what users frequently mention using word clouds (imported via Python visuals if needed).**
* **Check-in Behavior:**
  + **Examine most-visited times/days using Power BI time-series visuals.**
  + **Understand business popularity through check-in frequencies.**

**4. Dashboard Creation in Power BI**

* **Interactive Dashboards:**
  + **Create slicers and filters for city, category, star rating, and review volume.**
  + **Allow dynamic drill-downs by business, user, and timeframe.**
* **Key KPIs & Visuals:**
  + **Star rating distributions**
  + **Review count vs average rating**
  + **Top 10 businesses per category or city**
  + **Time-based trend charts for reviews/check-ins**

**Tools & Technologies:**

* **Programming Language: Python**
* **Libraries: pandas, NumPy, NLTK, Scikit-learn, Matplotlib, Seaborn, TextBlob, SpaCy, Transformers**
* **Visualization Tools: Power BI (for final dashboard), Plotly, Streamlit (for optional prototyping)**
* **IDE: Jupyter Notebook**
* **Cloud Platform: AWS (Amazon Web Services)**

**Storage: Amazon S3**

**Compute: Amazon EC2 / SageMaker**

* **Dataset Source: Yelp Open Dataset**

**Expected Outcomes:**

* Visual and statistical insights from Yelp data
* A recommendation engine suggesting personalized businesses
* A predictive model forecasting user ratings

**Conclusion:**

This project bridges data analytics with practical machine learning by using real-world data from Yelp. It demonstrates how intelligent systems can assist users in decision-making by offering personalized recommendations and predictions, which are widely used in platforms like Amazon, Netflix, and Yelp itself.