Build a Natural Language Chabot for Excel-Based Insights

Problem Statement

A mid-sized enterprise relies heavily on Excel reports to track all its business operations, including sales, inventory, employee performance and financial KPIs. Often, business users need quick insights from these files without waiting on the technical and reporting teams to filter, aggregate or visualize the data effectively.

As an intern at NeoStats, you are tasked with designing a conversational assistant that can read and understand Excel files and help users get the answers they need simply by asking questions in plain English.

The goal of this assignment is to evaluate your ability to work with structured data, integrate LLMs, reason through queries and build an intuitive user facing application.

Your Objective

You will build a web-based chat assistant that:

- Accepts an Excel file from the user
- Reads and analyses the data
- Allows users to ask natural language questions
- Returns answers as texts, tables or charts depending on the query.

Data Assumptions

The assistant should be able to handle general tabular data uploaded as an Excel file. Please keep the following expectations in mind:

1. **Excel File Format**
    - Input will be a standard .xlsx file.
    - The file will contain one sheet with a clearly labelled header row and structured data.
    - Expect up to 500 rows and 10-20 columns.

2. **Data Types**
    - The dataset will contain a mix of numeric columns, categorical columns and binary indicators (yes/no flags).
    - Some columns may contain missing or null values.
    - Columns names may include spaces, special character, or inconsistent casing.

3. **Assistant Behaviour**
    - Automatically infer and understand column types
    - Normalize column names internally for processing
    - Avoid relying on hardcoded column names or schemas.

4. **Types of Questions to Support**
   The assistant should support natural language queries like:
   - Statistically summaries (e.g. What is the average income?)
   - Filtered queries (e.g. How many customers are under 30?)
   - Comparisons or groupings (e.g. Compare loan defaults by gender)
   - Visual insights (e.g. Show a bar chart of transaction count by job)

5. **Visualizations**
   If the question implies a trend, comparison, or distribution, generate appropriate charts (e.g. bar chart, histograms, line chart). Ensure the charts are readable and labelled clearly.

6. **Important Notes**
   - Do not hardcode any column names, values or assumptions.
   - Focus on making your assistant flexible and schema agnostic.
   - Treat this as a general-purpose insight engine. Your goal is to make the assistant useful even if the Excel schema changes.
   - You may leverage the open-source LLMs or OpenAI as needed; however, please note that NeoStats will not be providing any APIs for this use case.

Submission Guidelines

1. **Host your application** Your conversational assistant should be available for us to test through a live hosted link.

   Suggested platforms:

   - Streamlit Cloud
   - Hugging Face Spaces

2. **Submit a GitHub Repository Link** Share a link to your public GitHub repository containing:

   - All project source code.
   - A clear and concise README.md

3. Please ensure to include a PPT document that describes your solution.

Evaluation Criteria:

- **Functional correctness:** The assistant correctly reads Excel files, processes data and answers questions accurately.
- **Reasoning and use of LLM:** Effective integration of the language model to understand queries and extract insights.
- **Chart generation and visualization:** Appropriate and clear charts are generated when questions require visual data representation.

- **Usability and interface design:** The app is easy to use, intuitive, and provides clear feedback or error messages.
- **Code quality and organization:** Code is clean, modular, well-documented, and easy to maintain.
- **Flexibility and generalization:** The assistant works well across different Excel schemas without hardcoded assumptions.
- **Creativity and bonus features (optional):** Additional useful features, thoughtful design improvements, or enhanced UX/UI elements.

Remember, this project is an opportunity to showcase your skills, creativity, and problem-solving abilities. Candidates are expected to showcase their analytical, logical thinking, and deduction skills throughout this challenge. Your approach to the problem, as well as your ability to present and defend your findings, will be central to the evaluation process.

Please ensure the work is your original work. Any kind of plagiarism or taking help from others will have adverse consequences. Refrain from using ChatGPT, Gemini or any other AI tool, this will lead to immediate disqualification. We strictly do not tolerate any kind of malpractice or misbehaviour.

Best of Luck!

Team NeoStats.