Large-Language Assisted Data Analysis Tool

Rijul Saini

Khoury College of Computer Sciences
Northeastern University
saini.ri@northeastern.edu

Abstract. In the rapidly evolving landscape of data analysis, there exists a significant gap between the availability of data and the ability of non-technical users to extract meaningful insights from it. Traditional data analysis tools often require advanced technical skills, limiting their accessibility and utility to a broader audience. Addressing this challenge requires the development of intuitive and user-friendly data analysis tools that leverage advanced technologies to facilitate data exploration and interpretation.
LLADA employs a combination of python programming and OpenAI's GPT-3.5 turbo API to provide users with a user-friendly interface for data analysis. By integrating natural language processing capabilities, the tool allows users to interact with datasets using plain language commands, eliminating the need for complex programming or query languages. The methodology involves summarizing the dataset, generating analysis goals, generating analysis results, and finally interpreting the results.
The key findings of the project include the successful development of a data analysis tool that empowers non-technical users to analyze and interpret datasets effectively. The tool's intuitive interface and language-assisted capabilities make it accessible to a wide range of users, enabling them to derive meaningful insights from data without requiring specialized technical expertise. Additionally, the project demonstrates the potential of advanced technologies such as natural language processing to democratize access to data analysis tools and promote data-driven decision-making across various domains.

Keywords: Large-Language models, OpenAI API, Data analysis, Python

1 Introduction

Data analysis plays a crucial role in various domains, including business, healthcare, finance, and academia. With the proliferation of data across industries, there is a growing need for advanced data analysis tools that can handle large and complex datasets efficiently. Recent advancements in machine learning and natural language processing have paved the way for the development of innovative approaches to data analysis, offering new opportunities to extract insights and drive decision-making.

The problem addressed by this project is the accessibility and usability of data analysis tools for non-technical users. Traditional data analysis tools often require advanced technical skills and knowledge of programming languages, making them inaccessible to individuals without a background in data science or computer programming. This poses a significant barrier to data-driven decision-making and hinders the democratization of data analysis capabilities across organizations and industries.

The main objectives of the project are to develop a user-friendly data analysis tool that:

- Empowers non-technical users to analyze and interpret complex datasets
- Provides an intuitive interface for interacting with data using natural language commands
- Leverages advanced technologies such as machine learning and natural language processing to automate data analysis tasks
- Facilitates informed decision-making and problem-solving based on data-driven insights

The project adopts a multi-faceted approach to address the problem, combining techniques from data science, machine learning, and natural language processing. The development process involves:

- Summarizing the dataset to generate a JSON summary for goal generation, along with a natural language summary.
- Analysis goals generation with some rationale for why the analysis is needed.
- Generation and execution of Python code for the analysis goals to obtain results.
- Generation of natural language interpretation of the analysis results for the users.

We hope to contribute to the data science community with the development of a user-friendly data analysis tool that bridges the gap between data availability and data literacy. The integration of advanced technologies such as machine learning and natural language processing automates data analysis tasks and enhances usability. The project promotes data-driven decision-making and problem-solving across various domains by democratizing access to data analysis capabilities, and aims to advance research and innovation in the field of data analysis.

2 Methodology

2.1 Problem Statement

This project tackles the challenge of making data analysis accessible to individuals without a technical background. Traditional data analysis tools often require programming skills and an understanding of statistical concepts, hindering their user base.

2.2 Significance

Democratizing data analysis holds immense potential across various sectors. Businesses can gain insights from customer data without relying on data scientists. Researchers from diverse backgrounds can analyze data relevant to their fields without needing to learn complex programming languages. Overall, this tool empowers individuals to leverage the power of data for informed decision-making.

2.3 Data Preparation

This tool is designed to be user-agnostic and function with any dataset a user uploads as a pandas data frame. There is no specific data collection step within the tool itself. Users are responsible for ensuring their data is clean and formatted appropriately for analysis.

2.4 Functionality

The core of this tool lies in the GPT-3.5-turbo LLM. GPT-3.5-turbo is a powerful language model trained on a massive dataset of text and code. This allows it to understand user instructions related to data analysis and generate Python code to achieve the desired outcome. The tool consists of four main functionalities: Summarizer, Goal generator, Analyzer and Interpreter.

1. Summarizer:
   - The tool reads the data and automatically generates a summary in JSON format.
   - The summary includes: Dataset name, Number of entries (rows) and data fields (columns), Data types for each column (e.g., integer, string, date), Basic statistics for numerical columns (e.g., mean, std dev.) and number of unique values for categorical columns.
   - Optionally, also generates a natural language summary of the dataset.
2. Goal generator:
   - Generates analysis goals from the data summary, also includes some rationale for why the analysis is needed.
   - Number of goals to be generated can be set by specifying the n argument (n = 3 by default).

  - The function returns a list of 'goals', where each 'goal' is a dictionary containing: index, question, analysis, and rationale.
3. Analyzer:
   - Generates and executes Python code for the analysis goals.
   - If multiple goals are passed as argument, all goals will be analyzed by default. Optionally, you can pass 'i' argument to specify any particular goal by index from goals, and only that analysis will be generated.
   - The generated analysis code's execution is output to the console, and the function returns a list of 'results' where each 'result' is a dictionary containing: index, question, analysis code, analysis result, and rationale.
4. Interpreter:
   - Generates natural language interpretation of the analysis results for the user.
   - The function returns and prints a string containing the questions and their results derived from the analyses.

The four modules are designed as a pipeline with the output of each module being fed as the input for the next module. Overall, the tool is designed to receive a data frame as input and produce a summary, analysis goals, analysis results, and result interpretation seamlessly without user intervention.

3 Future Work

3.1 Limitations of the Project

Despite successful implementation of the tool, the project has several limitations due to the limited time frame that warrant further investigation. These include:

1. Limited input scope: The project focuses primarily on structured datasets and may not fully address the needs of users working with unstructured or semi-structured data.
2. Limited analysis scope: The tool struggles to successfully execute complex analyses like multiple regression or clustering.
3. Performance constraints: The performance of the tool may be limited by computational resources, particularly when working with large datasets or complex analysis tasks.
4. User interface: While the tool aims to be user-friendly, there may be opportunities to further improve the user interface and user experience to enhance usability and accessibility.

3.2 Future Research Directions

To address these limitations and build upon the project's successes, several avenues for future research and development could be explored:

1. Expansion to unstructured data: Investigate techniques for analyzing unstructured and semi-structured data, such as text and image data, to broaden the tool's applicability and utility.
2. Expand analysis scope: Further refinements to the code generation algorithm could allow for more robust analyses to be performed.
3. Performance optimization: Explore strategies for improving the performance and scalability of the tool, such as parallel processing, prompt optimization, or even training LLMs on data specific to the project needs.
4. User interface enhancements: Adding front-end functionality for ease of use, like using click-and-drop instead of calling the python package.

The Large-Language Assisted Data Analysis Tool has the potential to have a significant impact on the field of data analysis by democratizing access to advanced analytical capabilities. By enabling non-technical users to harness the power of data, the tool can facilitate informed decision-making, drive innovation, and unlock new opportunities for data-driven insights across industries. Furthermore, the project contributes to ongoing efforts to bridge the gap between data availability and data literacy, thereby empowering individuals and organizations to leverage data effectively for informed decision-making and problem-solving. By addressing these areas of future work, the project can continue to evolve and become a powerful multi-purpose analysis tool.

4 Conclusions

In conclusion, the Large-Language Assisted Data Analysis Tool represents a significant advancement in the field of data analysis, providing a user-friendly platform for non-technical users to interact with and derive insights from complex datasets. By leveraging advanced technologies such as machine learning and natural language processing, the tool empowers users to make informed decisions and drive innovation across industries. The project's methodology, key findings, and future directions contribute to ongoing efforts to democratize access to data analysis capabilities and promote data-driven decision-making. As the project continues to evolve and expand, it holds the potential to revolutionize the way data is analyzed and interpreted, ultimately shaping the future of data science and machine learning.

5 Related Work

While there are many LLM-powered tools currently available that utilize their capabilities for data related tasks, there are no tools to directly analyze any given data without user intervention. Our tool fills a unique gap in the current data science landscape by providing an almost completely autonomous tool for data analysis. There were many projects and studies that contributed to the successful completion of this project, which are listed below:

1. Microsoft Lida - a tool for generating data visualizations by utilizing LLMs. https://microsoft.github.io/lida/
2. OpenAI API documentation for implementing GPT assistants. https://platform.openai.com/docs/overview
3. How LLMs and Data Analytics Work Together. https://www.pecan.ai/blog/llm-data-analytics-work-together/
4. Qualitative data analysis using LLMs. https://github.com/Gamma-Software/llm_qualitative_data_analysis
5. Build a Custom Langchain Tool for Generating and Executing Code. https://betterprogramming.pub/building-a-custom-langchain-tool-for-generating-executing-code-fa20a3c89cfd
6. Python packaging user guide. https://packaging.python.org/en/latest/tutorials/packaging-projects/