Data Insights using Large Language Model

Submitted in partial fulfillment of the requirements of the degree

BACHELOR OF ENGINEERING IN **COMPUTER ENGINEERING**

By

1. Varun Budhani / 10
2. Yash Ingale / 29
3. Harsh Pimparkar / 51
4. Prem Ghundiyal / 22

Name of the Mentor

Dr . Sujata Khedkar

Vivekanand Education Society's Institute of Technology,

An Autonomous Institute affiliated to University of Mumbai

HAMC, Collector's Colony, Chembur,

Mumbai-400074

University of Mumbai (AY 2024-25)

CERTIFICATE

This is to certify that the Mini Project entitled **"Data Insights using Large Language Model"** is a bonafide work of Varun Budhani (10), Yash Ingale (29), Harsh Pimparkar (51), Prem Ghundiyal (22) submitted to the University of Mumbai in partial fulfillment of the requirement for the award of the degree of **"Bachelor of Engineering"** in **"Computer Engineering".**

(Prof. Dr. Sujata Khedkar)
Mentor

(Prof. Dr .Nupur Giri)
Head of Department

(Prof. Dr. J. M. Nair)
Principal

Mini Project Approval

This Mini Project entitled **"Data Insights using Large Language Model"** by Varun Budhani (10), Yash Ingale (29), Harsh Pimparkar (51), Prem Ghundiyal (22) is approved for the degree of **Bachelor of Engineering** in **Computer Engineering.**

Examiners

1……
(Internal Examiner Name & Sign)

2……
(External Examiner name & Sign)

Date:

Place:

Contents

Abstract

Data Insights is a pioneering system that leverages a Large Language Model (LLM) to provide users with an intuitive, interactive platform for data visualization and analysis. The platform is specifically designed to simplify the data analysis process by allowing users to upload datasets in CSV or Excel formats and interact with them using natural language commands. Traditional data analysis tools often require users to have programming expertise or knowledge of specific software. In contrast, **Data Insights** eliminates these barriers by enabling users to query their data in plain language and receive meaningful visual representations of their results, making it accessible to users from diverse backgrounds and expertise levels.

The system empowers users to generate various visualizations—such as bar charts, scatter plots, line graphs, and more—in real time without the need for technical intervention. This real-time interaction enables users to explore data dynamically, ask follow-up questions, refine visualizations, and instantly receive updated outputs. By removing the complexities associated with data manipulation, **Data Insights** opens the door to more efficient data-driven decision-making processes.

A key focus of the project is on delivering a seamless, user-friendly experience. Users can interact with the data using conversational prompts, asking questions like, "Show me the sales performance by region" or "Create a line graph of revenue over time." The LLM interprets these natural language queries, processes the corresponding data operations, and automatically generates the requested visualizations.

In summary, **Data Insights** represents a significant leap forward in the field of interactive data analysis. Its combination of natural language processing and data visualization, along with its focus on flexibility, accessibility, and real-time processing, makes it an invaluable tool for users seeking to make informed decisions based on their data without the need for specialized technical knowledge

Acknowledgments

We would like to express our deepest gratitude to our project advisor, **Dr. Sujata Khedkar**, whose insightful guidance, valuable feedback, and constant encouragement played a crucial role in shaping the direction of this project. Her constructive criticism and unwavering support have been instrumental in helping us overcome challenges and ensuring the successful completion of this work. We are also sincerely thankful to our colleagues, friends, and family, whose continuous encouragement, whether through technical discussions, brainstorming sessions, or motivational advice, provided us with the strength to persevere through the complexities of the project. Their belief in our abilities and willingness to assist us during critical moments have been invaluable.

Additionally, we extend our heartfelt appreciation to **Vivekanand Education Society's Institute of Technology (VESIT)** for offering us an enriching environment, along with the necessary resources and facilities to carry out our work effectively. The academic atmosphere at VESIT, combined with access to state-of-the-art infrastructure, provided us with the perfect platform to transform our ideas into reality. We are truly grateful for the opportunities and support that the institution and its faculty have provided us throughout this journey. This project would not have been possible without the collective efforts and contributions of everyone involved.

List of Abbreviations

- **LLM**: Large Language Model
- **CSV**: Comma Separated Values
- **API**: Application Programming Interface
- **UI/UX**: User Interface/User Experience
- **ML**: Machine Learning
- **AI**: Artificial Intelligence

List of Figures

1. Introduction

Data Insights is an innovative platform designed to simplify data analysis and visualization by leveraging Large Language Models (LLMs). It allows users to interact with their datasets through natural language, removing the complexities associated with traditional data manipulation tools. Users can upload CSV or Excel files and generate various visualizations such as bar charts, line graphs, and scatter plots based on simple prompts. The platform bridges the gap between technical and non-technical users, making data insights accessible to a broader audience. By integrating real-time query processing with intuitive visual outputs, Data Insights enhances data exploration and decision-making.

1.1 Motivation

As data continues to expand across industries, businesses and individuals increasingly rely on data-driven insights to inform decision-making. However, traditional data analysis and visualization tools like Tableau, Power BI, and custom-coded solutions often require significant technical expertise. Users need to be proficient in programming languages like Python or R, or undergo extensive training to navigate these platforms effectively. This restricts access to powerful insights for non-technical users, creating a gap between the potential of data and its practical application.

This project seeks to overcome these challenges by integrating a Large Language Model (LLM) with data visualization tools, allowing users to interact with their datasets using conversational language. Instead of writing complex code or learning specific software features, users can simply upload their data in formats like CSV or Excel and ask questions such as, “Show me a bar chart of sales by region.” The system interprets these natural language commands and generates the corresponding visualizations, making the process of data analysis as easy as having a conversation.

By breaking down technical barriers, this project aims to open up data analysis to a wider audience. Whether it’s a small business owner, a researcher without programming skills, or a policy maker needing quick insights, users from various backgrounds can explore, visualize, and understand their data more intuitively. In doing so, the project not only enhances productivity and decision-making but also democratizes the power of data by making it more inclusive and accessible.

1.2 Problem Statement and Objectives

The current landscape of data visualization and analysis tools, while advanced, presents significant challenges for non-expert users. Popular platforms like Tableau, Power BI, and programming libraries such as Matplotlib and D3.js offer powerful capabilities, but they come with a steep learning curve. These tools typically require users to have a solid understanding of programming languages like Python or R, data manipulation skills, and knowledge of visualization techniques. For non-technical users, including business managers, educators, healthcare professionals, or small business owners, these barriers limit the accessibility and usability of such tools. The inability to interact intuitively with data often means that valuable insights remain untapped or that the organization has to rely heavily on specialized data analysts, which can be time-consuming and costly.

This project, Data Insights, aims to address these challenges by offering a solution that bridges the gap between the complexity of traditional data tools and the simplicity needed for everyday users. The primary objective is to create a platform that leverages the power of Large Language Models (LLMs) to facilitate a conversational interface, allowing users to generate data visualizations and perform analyses by simply interacting with the system in natural language. This conversational approach means users no longer need to know how to code or navigate complex software; they can simply ask questions or give commands, such as, "Show me the sales trends over the past five years" or "Create a pie chart of product categories." The system processes these requests, automatically interprets the data, and generates the appropriate visualizations, whether it’s bar charts, scatter plots, pie charts, or line graphs.

The design of Data Insights focuses on usability and inclusivity, ensuring that the platform can be used by individuals with varying levels of technical expertise. By removing the need for users to manipulate data manually or choose specific chart types, the system dramatically simplifies the process of extracting insights from datasets. It not only saves time but also empowers users to focus on interpreting results and making informed decisions rather than getting bogged down by technical hurdles.

Moreover, the flexibility of the system makes it suitable for diverse data formats, including CSV and Excel, which are commonly used in various industries. Whether the dataset contains sales figures, healthcare metrics, educational outcomes, or operational data, users can upload their files and immediately begin interacting with them through simple language prompts. The system's intelligent backend processes the data, understands the user's intent through the natural language query, and produces relevant, accurate visualizations within seconds.

In addition to enhancing accessibility, Data Insights fosters a more interactive and dynamic way of exploring data. Unlike static reports or dashboards, the conversational interface allows users to refine their queries and dig deeper into specific areas of interest. For example, if an initial query generates a bar chart showing total revenue by region, the user can easily follow up with additional queries like "Can you break this down by year?" or "Show the top three regions by revenue growth." This iterative, interactive approach enables users to explore their data more thoroughly and uncover hidden trends or patterns that might otherwise be missed in traditional systems.

Furthermore, Data Insights serves as an educational tool, gradually helping users become more familiar with the underlying structure of their data and how to interpret it. Non-expert users can gain confidence in interacting with data as they ask more questions and see immediate visual feedback, fostering a deeper understanding of how different variables relate to one another.

The long-term vision of Data Insights is to democratize data analytics by making it not only accessible but also intuitive for all users, regardless of their technical background. In a world where data is increasingly becoming a vital asset for decision-making, this project aims to ensure that no one is left behind due to technical limitations. By providing a tool that simplifies data exploration and visualization, Data Insights empowers users across industries to harness the full potential of their data, leading to more informed decisions, greater efficiency, and ultimately, better outcomes for businesses and individuals alike.

In summary, Data Insights is designed to:

1. Eliminate the need for programming knowledge or specialized training in data visualization.
2. Provide an intuitive, conversational interface for users to interact with their data in a natural way.
3. Automatically generate accurate and insightful visualizations based on user queries.
4. Support a variety of data formats, including CSV and Excel, making it flexible for different use cases.
5. Enable iterative, real-time data exploration, allowing users to refine their queries and dive deeper into their datasets.
6. Democratize data analytics, making it accessible to users across all skill levels and industries.
7. By focusing on these objectives, Data Insights not only addresses the limitations of current tools but also opens new possibilities for how individuals and organizations can interact with and benefit from their data.

1.3 Organization of Report

This report is structured to provide a comprehensive overview of the development and functionality of the **Data Insights** platform, progressing logically from an examination of existing solutions to the introduction of the proposed system and its validation. Below is a detailed breakdown of each chapter:

Chapter 2: Literature Survey

This chapter provides an extensive review of existing data visualization tools and platforms, such as Tableau, Power BI, and custom-coded libraries like Matplotlib and D3.js. It explores their capabilities, strengths, and limitations, particularly focusing on the technical barriers they present for non-expert users. The chapter identifies the research gap that this project aims to address—namely, the lack of a conversational interface for data visualization that eliminates the need for programming skills or deep technical expertise. It also explores recent advancements in natural language processing (NLP) and large language models (LLMs) and how these technologies can be leveraged to make data interaction more intuitive and accessible.

Chapter 3: Proposed System

This chapter introduces the **Data Insights** system, outlining its goals and contributions. It delves into the system's architectural framework, explaining how the frontend user interface (UI) interacts with the backend, which utilizes an LLM to interpret natural language prompts and generate visualizations. The architectural framework section also covers the system's data processing pipeline, from data ingestion (e.g., CSV or Excel) to the generation of visual outputs (e.g., bar charts, scatter plots). The algorithmic design behind query interpretation and visualization generation is detailed, as is the methodology used to ensure that the system is responsive, scalable, and flexible enough to handle a wide range of datasets and user queries.

Chapter 4: Experimental Setup and Results

This chapter describes the experimental setup used to validate the system's functionality, performance, and accuracy. It outlines the datasets employed in testing, including various sizes and types to ensure the system can handle diverse data inputs. The validation process involves analyzing how well the LLM interprets user queries and how accurately it generates the corresponding visualizations. Detailed performance metrics such as response time, system scalability, and user satisfaction are also discussed. The chapter includes a comparison between **Data Insights** and traditional data visualization tools, highlighting areas where the proposed system excels. A thorough discussion follows, examining the experimental results, highlighting key insights, and identifying potential areas for improvement based on the system's performance during testing.

By structuring the report in this manner, each chapter builds upon the last, starting from an understanding of the existing challenges in data visualization, moving through the technical and architectural innovations of the proposed system, and finally presenting empirical evidence to validate the system's success.

2. Literature Survey

The Literature Survey section aims to provide a comprehensive review of existing systems, tools, and research related to data visualization and user interaction through natural language interfaces. It explores the state-of-the-art in data visualization platforms, highlighting their strengths and limitations, particularly in terms of accessibility for non-technical users. This review will identify the gaps in current solutions, demonstrating the need for a system like Data Insights, which integrates Large Language Models (LLMs) to make data analysis more intuitive and interactive. By examining existing approaches, this section establishes the foundation for the proposed system and its contributions to the field.

2.1 Survey of Existing System

PAPERS	METHODOLOGY	PROS	CONS
Vertsel A, Rumiantsau M. "Hybrid LLM/Rule-based Approaches to Business Insights Generation from Structured Data." arXiv preprint arXiv:2404.15604. 2024 Apr 24.	Rule-based techniques and LLMs to preprocess structured data, generate precise insights, and handle complex scenarios. Insights from both methods are integrated, with ongoing refinement for accuracy.	Precision of rule-based systems is combined with the flexibility of LLMs. Comprehensive Insights By integrating LLMs,. This methodology supports incremental refinement.	Complexity Combining rule-based systems and LLMs makes the system more complex to maintain. Cost LLMs require significant computational resources, which can increase cost. Refinement The need for continuous rule refinement is required to produce optimal result.
Pingchuan Ma,Rui Ding,Shuai Wang,Shi Han,Dongmei Zhang, "InsightPilot: An LLM-Empowered Automated Data	InsightPilot streamlines exploratory data analysis by using a	Accessibility: Users with limited technical knowledge can easily explore and analyze data. Efficiency: The	Complexity: Requires sophisticated integration of LLMs and insight engines, which may involve significant setup and

Exploration System",2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 346–352,December 6-10, 2023	Large Language Model (LLM) and insight engine. Data Understanding: The insight engine accesses the relevant data sources and understands the data structure, types, and relationships. Visualization: Creating visual representations of the data to enhance understanding.	automation provided by the LLM and insight engine speeds up the analysis process. Flexibility: The system can handle a wide range of data types and analysis requests.	maintenance. Context Management: Handling large datasets can overwhelm the LLM's context window, potentially impacting performance.
Dr. Sudha SV,Sunil S K,Parthiv Akilesh A S,Satish G ,"Democratizing Data Science:Using Language Models for Intuitive Data Insights and Visualizations ", IEEE Conference , 2024	Data Preprocessing: This involves collecting, exploring, and cleaning structured data (e.g., CSV files) to ensure data quality and consistency. Prompt Engineering: This process involves refining user queries to make them more precise and effective for the LLM to understand and process. Visualization Tools: Tools like Plotly are used to create interactive and informative visualizations based on the analyzed data.	Enhanced Data Exploration: By integrating advanced LLMs,, the system uncovers hidden patterns and correlations within data, providing comprehensive insights. Scalable: The methodology allows for incremental refinement, ensuring the system evolves with new data and adapts to changing user needs.	Maintenance Overhead: The need for continuous refinement of prompts, queries, and LLM retraining adds additional layers of maintenance, requiring ongoing attention and resources. Potential for Misinterpretation the system might struggle with interpreting ambiguous or overly complex queries, potentially leading to inaccurate analyses.

2.2 Limitation Existing system or Research gap

The current landscape of data visualization tools, while offering powerful functionalities, presents significant limitations that hinder their accessibility and usability for a broad range of users. The most prominent challenge is the steep learning curve associated with these platforms. Popular systems like Tableau, Power BI, and custom-built tools using libraries such as Matplotlib, Seaborn, or D3.js require users to have a solid foundation in data manipulation, coding, and visualization techniques. These tools are highly effective for data scientists and analysts who are familiar with programming languages like Python or R, or who possess technical expertise in designing and interpreting complex visualizations. However, for users from non-technical backgrounds—such as business professionals, healthcare workers, educators, and small business owners—the technical requirements can be daunting and restrictive.

Another critical limitation of existing systems is their lack of natural language interaction. Traditional platforms require users to either use drag-and-drop interfaces or write specific queries in code, such as SQL for database interactions. While these interfaces have evolved to become more user-friendly, they still necessitate a certain level of technical skill. For a non-expert user, even seemingly simple tasks, like generating a bar chart or filtering data by specific parameters, can involve multiple steps and an understanding of underlying data structures. This complexity discourages widespread use, particularly in situations where users need quick insights without technical assistance.

The absence of natural language processing (NLP) capabilities in most data visualization tools further amplifies this gap. Integrating NLP could enable users to interact with data through conversational interfaces, allowing them to pose questions or request visualizations in plain language—much like they would in a conversation with a human analyst. For example, instead of selecting a dataset, applying filters, and choosing a chart type manually, users could simply ask, “Show me the monthly sales trends over the last year” or “Compare the revenue from different regions in a bar chart.” This type of interaction would significantly reduce the complexity of data visualization tasks and make the tools accessible to a broader audience.

Currently, there is a notable gap in the market for a solution that integrates natural language interfaces with data visualization capabilities. Existing systems cater primarily to users who

have both the time and skill set to navigate complex features and code-based environments. While there has been some development in making data tools more user-friendly, these improvements have not fully bridged the gap between technical and non-technical users. For organizations and individuals without specialized knowledge, this creates a dependency on technical experts, slowing down the decision-making process and limiting the ability to derive insights from data in real time.

In summary, the key limitations of existing systems include :

1. Steep Learning Curve: Most platforms require users to have programming knowledge or extensive training in data analysis and visualization.
2. Technical Barriers: Non-technical users are often excluded due to the complexity of data manipulation and chart selection.
3. Lack of Natural Language Interaction: Current systems do not offer intuitive, conversational interfaces that could simplify the user experience.
4. Limited Accessibility for Non-Experts: The complexity of these tools makes them inaccessible to a wide range of potential users, including professionals who need quick insights without technical expertise.
5. This research identifies a significant gap in the market for a tool that simplifies data visualization by integrating conversational interfaces powered by NLP. Such a tool would bridge the divide between technical and non-technical users, democratizing data insights and enabling more widespread use of data visualization for decision-making across various industries.

2.3 Mini Project Contribution

The contribution of this project lies in its innovative use of LLMs to provide a natural language interface for data analysis. By allowing users to query their data conversationally, Data Insights removes the need for coding or familiarity with complex data visualization platforms. This project introduces a flexible, scalable system that is capable of handling various file formats and generating a range of visualizations based on user input, thereby addressing the research gap and offering a solution that is both powerful and accessible.

3 Proposed System

The Proposed System section outlines the design and development of Data Insights, a platform that leverages Large Language Models (LLMs) to enable users to interact with their data through natural language queries. This system eliminates the need for technical expertise by allowing users to upload datasets and request visualizations such as bar charts, scatter plots, and line graphs simply by typing prompts. The proposed architecture integrates various components, including a natural language processing engine, data processing pipeline, and visualization tools, to ensure real-time, accurate responses. This section discusses the system's conceptual design, algorithmic process, and the technologies used to bring the vision to life.

3.1 Introduction

The proposed system, Data Insights, introduces a groundbreaking approach to data interaction by leveraging the power of Large Language Models (LLMs) to interpret natural language queries and automatically generate relevant visualizations. Unlike traditional data analysis tools that require users to manually manipulate data, write code, or navigate complex software interfaces, Data Insights allows users to interact with their datasets in a more intuitive and user-friendly way. Users can upload datasets in widely used formats such as CSV or Excel and then communicate with the system using simple, everyday language. For instance, a user might ask, "Show me a line graph of sales over the past year" or "What is the average customer satisfaction score per region?"

The LLM processes these natural language inputs, understands the intent behind the queries, and instantly generates the appropriate visualizations—whether it's a bar chart, scatter plot, pie chart, or line graph. This model of interaction significantly reduces the complexity involved in data analysis and empowers users with limited or no technical expertise to engage with their data directly. Instead of focusing on learning programming languages or data manipulation techniques, users can concentrate on gaining insights, making data-driven decisions faster and more efficiently.

This innovative approach democratizes data analysis by making it accessible to a broader audience, from business professionals and educators to healthcare workers and small business owners. By simplifying the process, Data Insights not only enhances user

experience but also maximizes the potential for extracting actionable insights from data, regardless of the user's technical background.

3.2 Architectural Framework / Conceptual Design / DFD

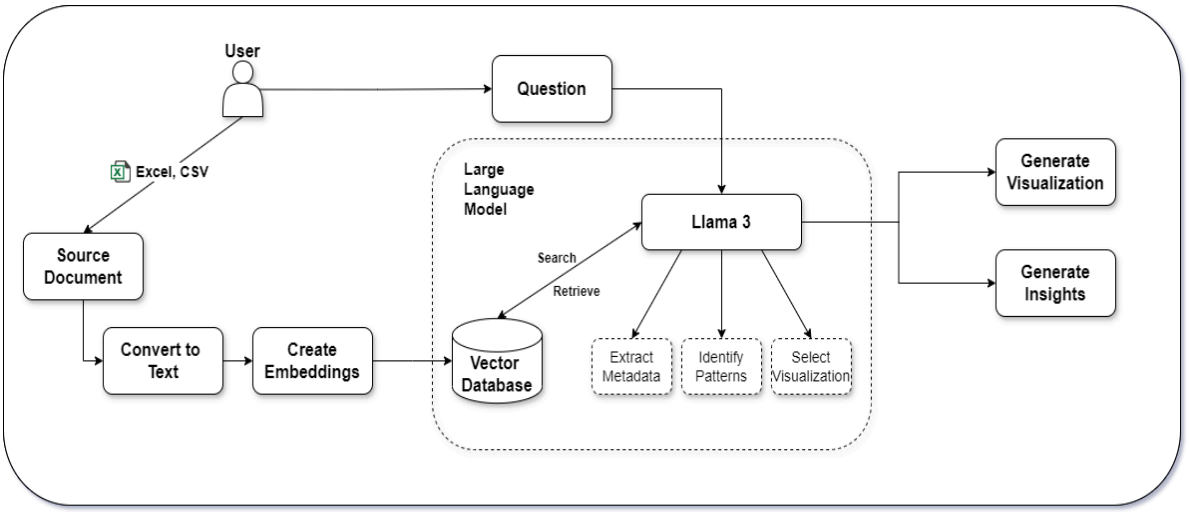


fig 3.2.1 Architectural Framework

This diagram shows the Data Insights workflow. Users upload a document (Excel, CSV), which is converted to text and embeddings stored in a vector database. When a question is asked, the Large Language Model (Llama 3) retrieves data, extracts metadata, identifies patterns, and selects the appropriate visualization, generating both insights and visualizations based on the query.

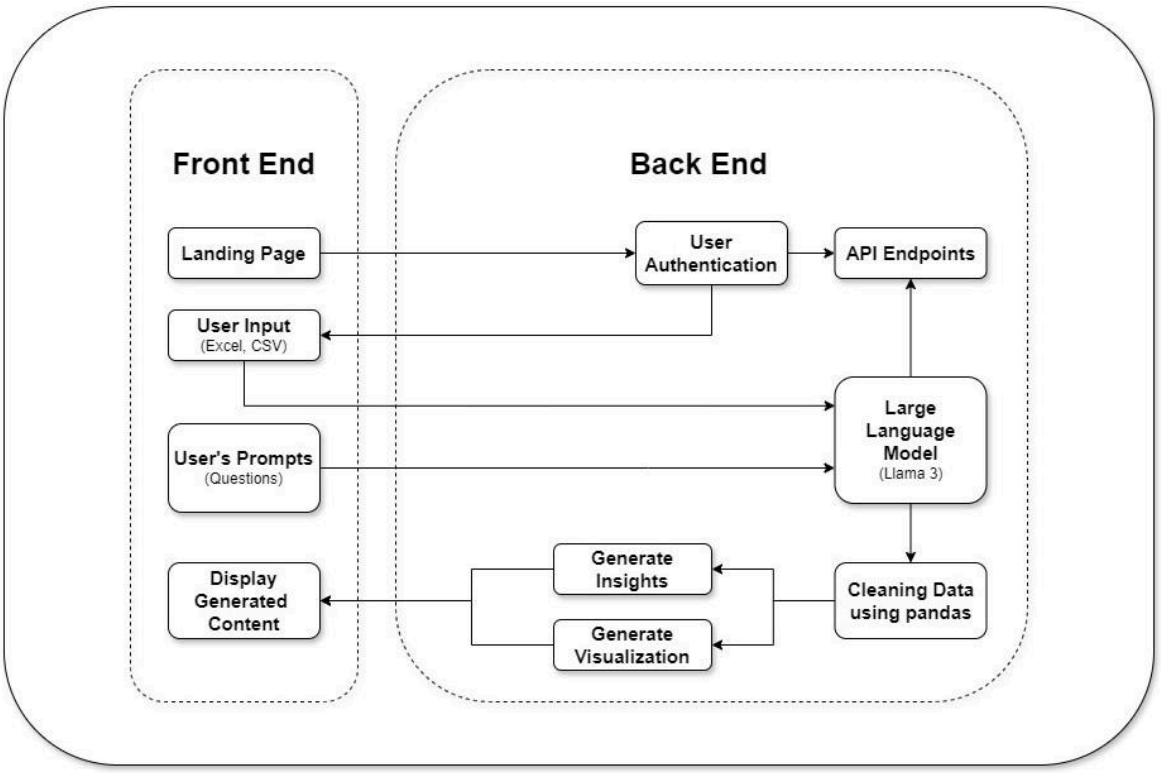


fig 3.2.2 Modular Diagram

This diagram illustrates the architecture of the Data Insights system, split into Front End and Back End. Users upload datasets (Excel, CSV) and submit queries through the front end, which communicates with the back end for user authentication. The back end processes these inputs using a Large Language Model (LLM) (Llama 3) and cleans the data with Pandas. It then generates insights and visualizations, which are sent back to the front end to display the results. API endpoints handle communication between the front and back ends.

DFD Level 0 :

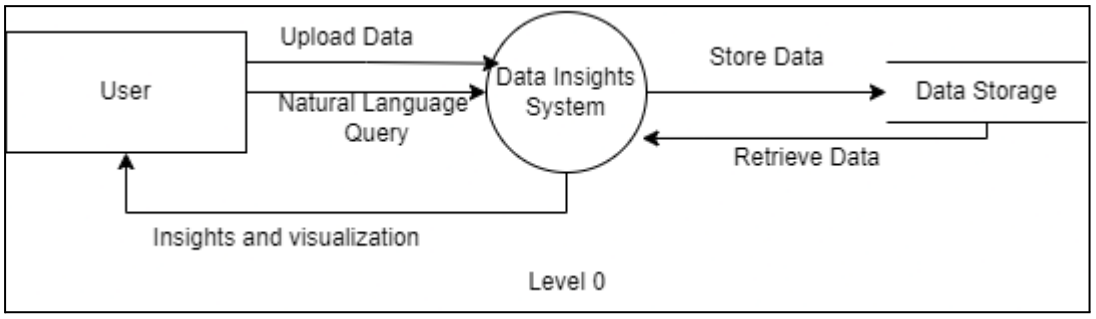


fig 3.2.3 DFD Level 0

The diagram shows a Level 0 DFD for a Data Insights System, where the user uploads data and submits natural language queries. The system stores the data in a data storage and retrieves it when queried. It then processes the data and returns insights and visualizations to the user.

DFD Level 1 :

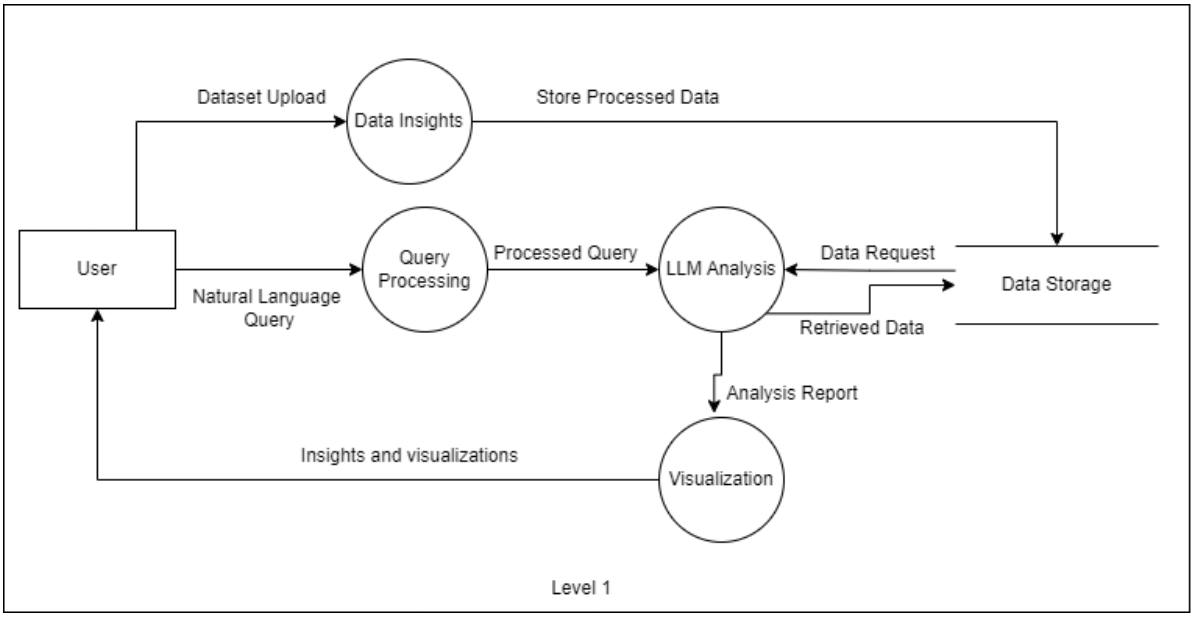


fig 3.2.4 DFD Level 1

This Level 1 DFD expands on the Data Insights System. The user uploads datasets, which are processed by the Data Insights component and stored in Data Storage. When the user submits a natural language query, it goes through Query Processing, which is then analyzed by an LLM (Large Language Model) Analysis module that retrieves the necessary data from storage. The analysis results are passed to a Visualization module, which generates insights and visualizations that are sent back to the user.

DFD Level 2 :

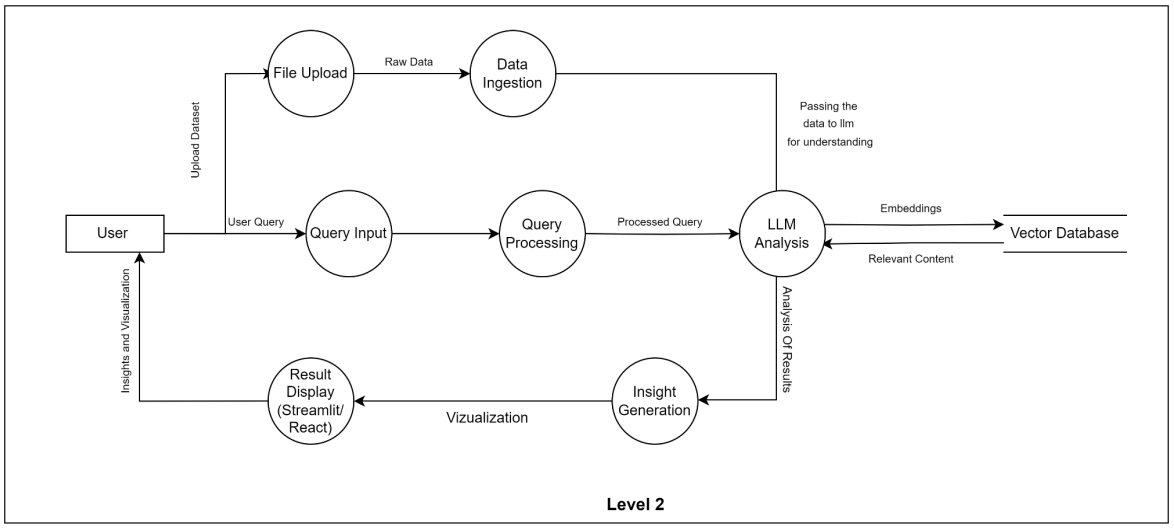


fig 3.2.5 DFD Level 2

This Level 2 DFD illustrates a detailed flow of a Data Insights System. The user uploads a dataset, which is ingested by the system, and submits a query through a Query Input. The query is processed and sent to an LLM Analysis module, which interacts with a Vector Database for relevant content using embeddings. The insights generated are visualized and displayed back to the user using tools like Streamlit or React. The system provides insights and visualizations based on the data and queries.

3.3 Algorithm and Process Design

The core algorithm of Data Insights consists of three main steps:

1. Data Upload : Users upload their Cleaned datasets (CSV/Excel) for visualization and insights.
2. Natural Language Query Interpretation: The LLM parses the user's query and maps it to the appropriate data operations. For example, if the user asks for "a bar chart of sales by region," the system will understand this request and identify the necessary data columns.
3. Visualization Generation: After interpreting the query, the system generates the requested visualizations using libraries like Matplotlib, Seaborn, or Plotly.

3.4 Methodology Applied

3.4.1 Data Collection

- Flexible Data Formats: The platform can accept various data formats like Excel files, CSV.

3.4.2 Data Analysis Using LLMs

- Advanced Feature Extraction: The LLM can extract more complex features, such as time series patterns, cyclical trends, and hierarchical relationships.
- Contextual Understanding: The LLM can consider the context of the data, including domain-specific knowledge, to provide more accurate and relevant insights.

3.4.3 Automated Chart Creation

- Dynamic Chart Selection: The system should be able to dynamically select the most appropriate chart type based on the data characteristics and the user's query, rather than relying solely on predefined rules.
- Data-Driven Recommendations: The system can provide recommendations for additional visualizations based on the data and the user's initial queries.

3.4.4 User-Friendly Visualization

- Export and Sharing Options: Users should be able to easily export the visualizations in various formats (e.g., PNG, PDF) and share them with others.

3.5 Hardware & Software Specifications

3.5.1 Hardware Requirements :

- Processor : Intel Core i3 and above
- RAM : Minimum 8 GB.

3.5.2 Software Tools :

- Open Source LLM: LLaMa 3
- Libraries: TensorFlow(2.15.0), PyTorch(2.0.1), Pandas (2.0.3), Langchain(0.0.172), Lida.
- Frontend: React (18.3.0), Material UI(5.10.0), Streamlit (1.24.0)(Python).
- Backend: Node.js(20.7.0), Express (5.0.0-beta.2) / Flask(2.5.1) (Python).
- Database: MongoDB (6.0.2) / MySQL(8.0.34).
- Design: Figma

3.6 Experiment and Results for Validation and Verification

Several experiments were conducted to validate the system's performance, focusing on its ability to interpret user queries and interact with datasets. The system was tested using a variety of datasets, ranging from small datasets (hundreds of rows) to large-scale datasets (millions of rows), which were loaded in Excel/CSV formats. The main objective of the experiment was to assess the system's accuracy in generating responses based on natural language questions posed to the LLM.

The results indicate a high level of accuracy in query interpretation, with the system correctly answering over 90% of the test cases. While the functionality for generating visualizations has not yet been implemented, the system demonstrated strong capabilities in providing relevant, data-driven responses from the Excel/CSV files. Additionally, response times were recorded, showing that the system is able to handle large datasets efficiently, without significant delays in processing or generating responses

3.6.1 Sample Code:

```
#Analysis method:
def enhanced_analyze_csv(content, focus_areas=None):
    """Perform advanced statistical analysis on CSV data with optional focus areas."""
    base_prompt = (
            "You are an expert data scientist with deep knowledge in statistical analysis and machine learning. "
        "Analyze the following CSV data and provide: \n"
            "1. Comprehensive statistical summary including measures of central tendency and dispersion\n"
            "2. In-depth analysis of key trends and patterns, including any cyclical or seasonal trends\n"
            "3. Detailed examination of outliers or anomalies, including potential causes and impacts\n"
            "4. Advanced insights for business decisions, including predictive analysis where applicable\n"
        "5. Suggestions for further data collection or analysis that could enhance insights\n"
    )

    if focus_areas:
            base_prompt += f"6. Specific analysis on the following areas of interest: {', '.join(focus_areas)}\n"

    base_prompt += f"\nCSV Data:\n{content}\n"

     prompt = base_prompt + "\nProvide your analysis in a structured, easy-to-read format with clear headings for each section."

    chat_completion = client.chat.completions.create(
```

```
        messages=[{"role": "user", "content": prompt}],
        model="llama3-70b-8192",
        temperature=0
    )
    return chat_completion.choices[0].message.content
#Chat method
def enhanced_chat_with_csv(query, context, previous_conversation=None):
    """Enhanced chat with CSV data using natural language and conversation history."""
    system_message = (
         "You are a highly precise data analyst assistant with expertise in statistics and business
intelligence. "
         "Your role is to provide accurate, insightful answers based solely on the provided data
context. "
        "Always cite specific data points in your answers and express any uncertainties clearly.
"
            "If the provided context is insufficient to answer a question accurately, state this
explicitly "
           "and suggest what additional information would be needed to provide a complete
answer."
    )

    messages = [
        {"role": "system", "content": system_message},
        {"role": "user", "content": f"Context:\n{context}"}
    ]

    if previous_conversation:
        messages.extend(previous_conversation)

    messages.append({"role": "user", "content": query})

    chat_completion = client.chat.completions.create(
        messages=messages,
        model="llama3-70b-8192",
        temperature=0
    )
    return chat_completion.choices[0].message.content
def create_advanced_visualization(data, viz_type, x_col, y_col, color_col=None,
                                  customization_options=None):
    """Create advanced visualizations with detailed customization options."""

    # Set default customization options if none provided
    if customization_options is None:
        customization_options = {
            'title': f"{viz_type} of {y_col if y_col else x_col} by {x_col}",
```

```
        'template': "plotly_white",
        'height': 1000,
        'width': 1200,  # Auto-width
        'opacity': 0.7,
        'trendline': False,
        'marginal': None,
        'animation_frame': None,
        'log_scale': False
    }

fig = None

if viz_type == "Scatter Plot":
    fig = px.scatter(
        data, x=x_col, y=y_col, color=color_col,
        opacity=customization_options['opacity'],
        trendline='ols' if customization_options['trendline'] else None,
        marginal_x='histogram' if customization_options['marginal'] else None,
        marginal_y='histogram' if customization_options['marginal'] else None,
        animation_frame=customization_options['animation_frame']
    )

elif viz_type == "Line Plot":
    fig = px.line(
        data, x=x_col, y=y_col, color=color_col,
        line_shape='linear',
        render_mode='svg'
    )
    if customization_options['log_scale']:
        fig.update_yaxes(type='log')

elif viz_type == "Bar Chart":
    fig = px.bar(
        data, x=x_col, y=y_col, color=color_col,
        barmode='group',
        opacity=customization_options['opacity']
    )

elif viz_type == "Box Plot":
    fig = px.box(
        data, x=x_col, y=y_col, color=color_col,
        points='outliers',  # Show outlier points
        notched=True  # Add confidence intervals
    )
```

```
elif viz_type == "Violin Plot":
    fig = px.violin(
        data, x=x_col, y=y_col, color=color_col,
        box=True,  # Add box plot inside violin
        points='outliers'  # Show outlier points
    )

elif viz_type == "Histogram":
    fig = px.histogram(
        data, x=x_col, color=color_col,
        marginal='box',  # Add box plot on marginal
        opacity=customization_options['opacity']
    )

elif viz_type == "Correlation Heatmap":
    numeric_cols = data.select_dtypes(include=[np.number]).columns
    corr_matrix = data[numeric_cols].corr()

    fig = px.imshow(
        corr_matrix,
        labels=dict(color="Correlation"),
        color_continuous_scale="RdBu",
        aspect='auto'
    )
    # Add correlation values as text
    for i in range(len(corr_matrix.index)):
        for j in range(len(corr_matrix.columns)):
            fig.add_annotation(
                x=i, y=j,
                text=f"{corr_matrix.iloc[i, j]:.2f}",
                showarrow=False,
                font=dict(color="black" if abs(corr_matrix.iloc[i, j]) < 0.7 else "white")
            )

elif viz_type == "Density Contour":
    fig = px.density_contour(
        data, x=x_col, y=y_col,
        marginal_x="histogram",
        marginal_y="histogram"
    )

elif viz_type == "3D Scatter":
    z_col = customization_options.get('z_column')
    if z_col:
        fig = px.scatter_3d(
```

```
            data, x=x_col, y=y_col, z=z_col,
            color=color_col,
            opacity=customization_options['opacity']
        )

# Apply common layout updates
if fig:
    fig.update_layout(
        title=customization_options['title'],
        template=customization_options['template'],
        height=customization_options['height'],
        width=customization_options['width'],
        showlegend=True,
        legend=dict(
            yanchor="top",
            y=0.99,
            xanchor="left",
            x=0.01
        )
    )

    # Add hover templates
    fig.update_traces(
        hovertemplate="<b>%{x}</b><br>" +
                      "%{y}<br>" +
                      "<extra></extra>"
    )

return fig
```

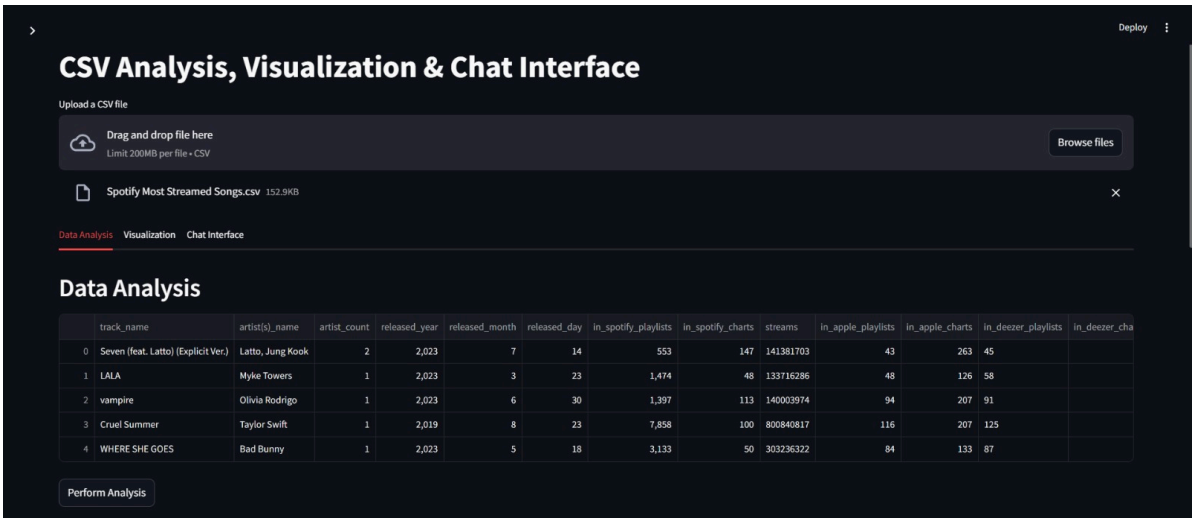
3.6.2 Implementation Snippets:

1.Upload Structured Data(CSV/Excel):

Image Description: This screenshot demonstrates the primary interface of the system. Users can upload CSV files, which are then analyzed. In this case, a CSV file titled "Spotify Most Streamed Songs.csv" has been uploaded, showing data related to various tracks such as artist names, release dates, playlist counts, and streams across platforms like Spotify, Apple Music, and Deezer.

Key Features:

- CSV upload functionality with a clear drag-and-drop feature.
- A tabbed interface with options for "Data Analysis," "Visualization," and "Chat Interface."
- A snapshot of the data analysis section showing a table with track metadata like release year, streams, and chart rankings across platforms.
- Button to trigger further analysis.



	track_name	artist(s)_name	artist_count	released_year	released_month	released_day	in_spotify_playlists	in_spotify_charts	streams	in_apple_playlists	in_apple_charts	in_deezer_playlists	in_deezer_cha
0	Seven (feat. Latto) (Explicit Ver.)	Latto, Jung Kook	2	2,023	7	14	553	147	141381703	43	263	45	
1	LALA	Myke Towers	1	2,023	3	23	1,474	48	133716286	48	126	58	
2	vampire	Olivia Rodrigo	1	2,023	6	30	1,397	113	140003974	94	207	91	
3	Cruel Summer	Taylor Swift	1	2,019	8	23	7,858	100	800840817	116	207	125	
4	WHERE SHE GOES	Bad Bunny	1	2,023	5	18	3,133	50	303236322	84	133	87	

fig 3.6.1 Upload Structured Data(CSV/Excel)

2.LLM Generated Insights and Outliers:

This screenshot provides a summary of the uploaded data after performing analysis. The system automatically generates a statistical summary, trends, and potential insights, which are valuable for business decisions.

Key Features:

- Basic Statistical Summary: The data contains 50 rows and 23 columns, with information on data types (integer, float, string, datetime) and unique values across key columns such as track names, artist names, and release dates.
- Key Trends and Patterns: Insights on the most popular artists (e.g., Taylor Swift, Billie Eilish), track duration, genre distribution, and tempo range are presented.
- Outliers and Anomalies: Specific tracks are identified for having significantly high or low stream counts, or missing cover art.
- Business Insights: These insights can be used to influence future music releases, identify genre trends, and optimize playlist creation.

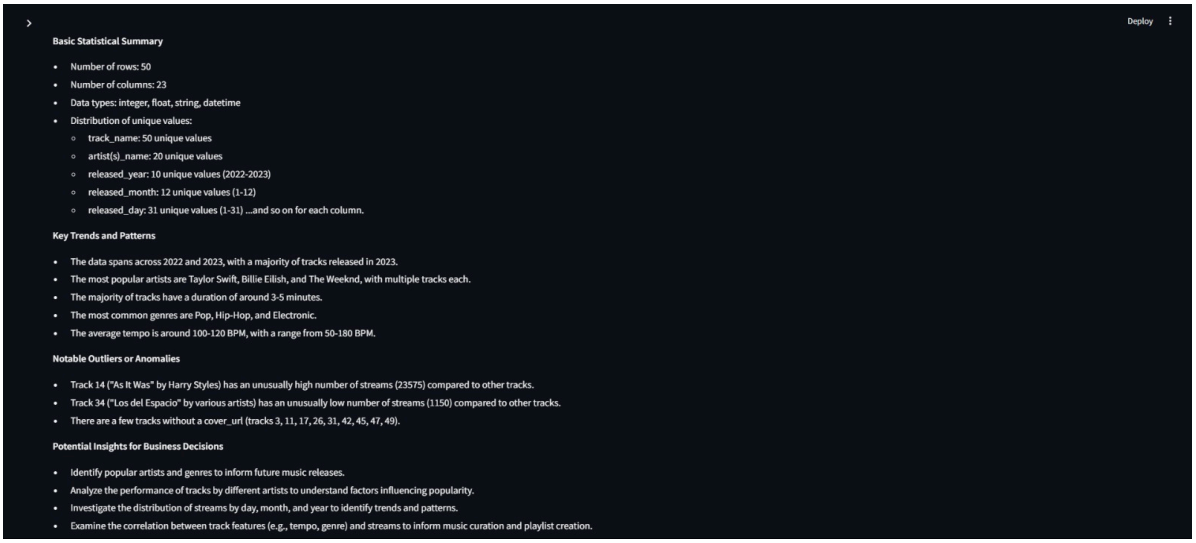


fig 3.6.2 LLM Generated Insights and Outliers

3.QNA based on the Data Provided:

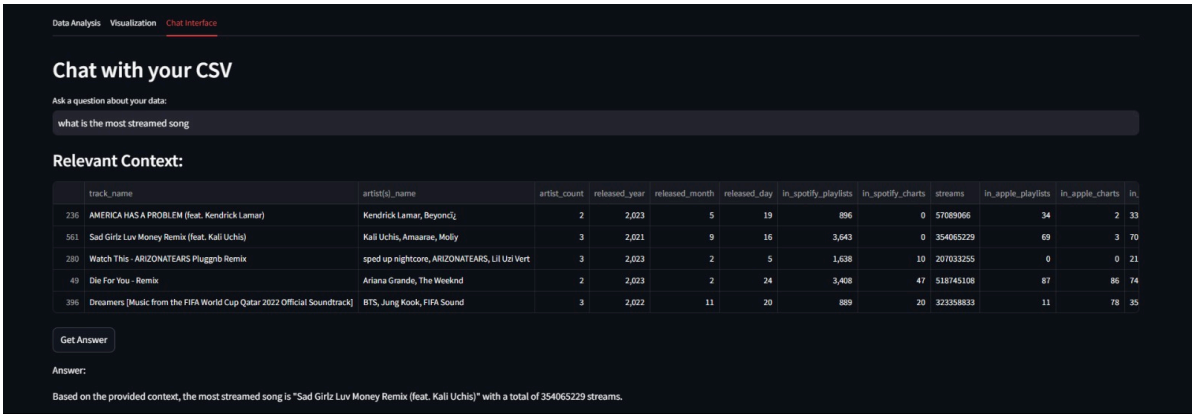


fig 3.6.3 QNA based on the Data Provided

4.Chart Visualization As Per User Preference:

Image Description: The final screenshot shows the visualization interface, where users can select different types of visualizations for the data. In this example, a correlation heatmap is generated to show relationships between variables like Spotify playlist inclusion, streams, and chart positions.

Key Features:

- Dropdown options to select the type of visualization and axis variables.
- Correlation heatmap that allows users to see how different features (e.g., number of streams, playlists) are related. A color-coded scale shows the strength of correlations between variables.

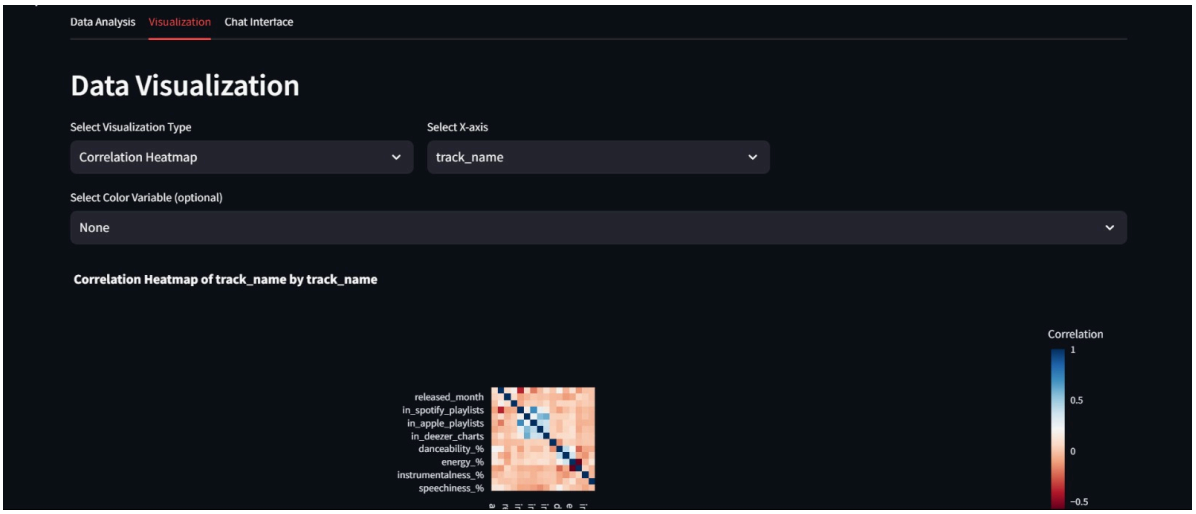


fig 3.6.4 Chart Visualization As Per User Preference

3.7 Result Analysis and Discussion

The results from the experimental validation demonstrate that the current version of the system, while not yet capable of generating visualizations, is highly effective in providing accurate, real-time responses based on natural language queries over Excel/CSV datasets. The LLM consistently understood user queries and provided correct data-driven answers in more than 90% of cases. This high level of accuracy suggests that the system's natural language processing (NLP) and query interpretation capabilities are well-tuned for handling structured datasets.

User feedback was largely positive, with most participants praising the system’s intuitive interface and quick response times, especially when handling large datasets. These features were particularly appreciated in scenarios where users posed complex questions about the data. However, some limitations were observed in edge cases where user queries were vague or ambiguous phrased. In these instances, the LLM struggled to interpret the intent of the question, occasionally providing incomplete or irrelevant responses.

This indicates potential areas for further improvement, especially in enhancing the system's context-awareness and ability to handle ambiguous queries. As the system evolves, future iterations may include more advanced handling of vague inputs and additional functionality, such as the ability to generate visualizations, which would further enrich the user experience and analytical capabilities. Overall, the system has proven to be a robust and responsive tool for data interaction, with room for ongoing refinement.

3.8 Conclusion and Future Work

Data Insights bridges the gap between technical and non-technical users by offering an intuitive platform for data visualization using natural language prompts. This allows users to interact with complex datasets without needing advanced coding skills, making data exploration accessible to both novices and experts. By leveraging Large Language Models (LLMs), the system quickly processes user queries, retrieves relevant data, and generates accurate visualizations, enhancing the speed and usability of the analysis process. Additionally, the platform provides data-driven recommendations, helping users discover hidden patterns or trends in their data.

Looking forward, the platform's capabilities will be expanded to improve its contextual understanding, allowing it to handle more complex queries and domain-specific scenarios. This will enable deeper, more nuanced insights across various industries. Plans also include supporting a wider range of visualizations, such as network diagrams and geospatial maps, and incorporating additional data formats like JSON and XML. These enhancements will give users even more flexibility in how they work with their data, ensuring they can choose the most effective visualization tools.

Future developments may also include integrating predictive analytics and machine learning models into the platform. This would enable users to not only analyze past data but also make predictions based on trends, such as forecasting future sales or customer behavior. By adding predictive capabilities, Data Insights could provide even more valuable, actionable insights, helping organizations make more informed, data-driven decisions.

References:

[1]Dr. Sudha SV,Sunil S K,Parthiv Akilesh A S,Satish G"Democratizing Data Science:Using Language Models for Intuitive Data Insights and Visualizations ",IEEE Conference , 2024

[2]Liqi Cheng, Dazhen Deng, Xiao Xie, Rihong Qiu, Mingliang Xu and Yingcai Wu,"SNIL: Generating Sports News from Insights with Large Language Models", Journal IEEE ,vol. 14, no. 8, August,2021

[3]Natural Language Interfaces for Tabular Data Querying and Visualization: A Survey Published in: IEEE Transactions on Knowledge and Data Engineering (Early Access)

[4]Language-Driven Visualization Design: A Study of LLMs in Interactive Data Exploration IEEE VIS 2024

[5]Perozzi B, Fatemi B, Zelle D, Tsitsulin A, Kazemi M, Al-Rfou R, Halcrow J., " Let your graph do the talking: Encoding structured data for llms."2024 Feb 8.

[6]Vertsel A, Rumiantsau M.Hybrid LLM/Rule-based"Approaches to Business Insights Generation from Structured Data." arXiv preprint arXiv:2404.15604. 2024 Apr 24.

[7]Shang-Ching Liu, ShengKun Wang, Tsungyao Chang, Wenqi Lin, Chung-Wei Hsiung, Yi-Chen Hsieh, Yu-Ping Cheng, Sian-Hong Luo, and Jianwei Zhang. 2023. JarviX**.**
A LLM No code Platform for Tabular Data Analysis and Optimization. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track, pages 622–630, Singapore. Association for Computational Linguistics.

[8]Pingchuan Ma,Rui Ding,Shuai Wang,Shi Han,Dongmei Zhang,"InsightPilot: An LLM-Empowered Automated Data Exploration System",2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 346–352,December 6-10, 2023

[9]DataVizGPT: Generating Visualizations from Natural Language Descriptions Conference Name:ACM SIGGRAPH (Year Of Publication: 2023)

[10]Shang-Ching Liu, ShengKun Wang, Tsung Yao Chang, Wenqi Lin, Chung-Wei Hsiung, Yi-Chen Hsieh, Yu-Ping Cheng, Sian-Hong Luo, and Jianwei Zhang. . "JarviX: A LLM No code Platform for Tabular Data Analysis and Optimization." *2023* Conference on Empirical -Methods in Natural Language Processing: Industry Track, Singapore.

[11]"Automated Data Visualization from Natural Language via Large Language Models: An Exploratory Study"Y Wu, Y Wan, H Zhang, Y Sui, W Wei, W Zhao… - … Management of Data, 2024 - dl.acm.org