# Data Insights Using LLM's

Mentor,
Dr.Sujata Khedkar
Associate Professor
dept name :- Computer Engineering

1st Varun Budhani
*dept. name:- CMPN*
*Vivekanand Education Society's Institute Of Technology*
Chembur,Mumbai,India
2022.varun.budhani@ves.ac.in

2nd Yash Inagle
*dept. name:- CMPN*
*Vivekanand Education Society's Institute Of Technology*
Chembur,Mumbai,India
2022.yash.ingale@ves.ac.in

3rd Harsh Pimparkar
*dept. name:- CMPN*
*Vivekanand Education Society's Institute Of Technology*
Chembur,Mumbai,India
2022.harsh.pimparkar@ves.ac.in

4rdPrem Ghundiyal
*dept. name:- CMPN*
*Vivekanand Education Society's Institute Of Technology*
Chembur,Mumbai,India
2022.prem.ghundiyal@ves.ac.in

***Abstract*--In an era where data is generated at an unprecedented rate, the ability to extract meaningful insights from complex datasets is essential for effective decision-making. This paper introduces a Data Insights and Visualization application that harnesses the power of Large Language Models (LLMs) to facilitate seamless user interaction with CSV and Excel files. The application empowers users to upload, clean, and analyze their datasets, enabling them to generate customized visualizations based on their specific queries. By integrating natural language processing (NLP) techniques, the system allows users to explore data intuitively, eliminating the need for advanced technical expertise in data analysis.**

***Our approach not only democratizes access to data analytics but also enhances usability and efficiency by providing an intuitive interface for discovering trends, correlations, and patterns in raw datasets. Unlike traditional data analysis tools that require manual data manipulation and complex scripting, our solution streamlines the process, reducing time and effort while improving accuracy and interpretability.**

***To assess the application's effectiveness, we conduct a comprehensive evaluation based on user feedback and performance metrics, demonstrating its ability to serve as a valuable tool for researchers, analysts, and business professionals. By bridging the gap between technical and non-technical users, our Data Insights and Visualization application aims to redefine the way individuals interact with data, fostering a more data-driven and informed decision-making process across various domains.**

***Index Terms*: data insights, data visualization, Large Language Model, CSV, Excel, natural language processing, data cleaning, user interaction, decision-making, trend analysis, data exploration, automated analytics, user feedback, performance metrics.**

## I. INTRODUCTION

In today's data-driven world, organizations across industries such as finance, healthcare, and education increasingly rely on data to drive decision-making and operational strategies. However, as the volume of data grows exponentially, extracting meaningful insights while managing complex datasets has become a significant challenge. The sheer scale of available information presents immense opportunities for improved efficiency and innovation but also increases the risk of data overload, making it difficult for users to derive actionable intelligence.

Recognizing these challenges, this project aims to develop an AI-powered Data Insights and Visualization application designed to simplify and enhance the data analysis process. By integrating advanced Large Language Model (LLM) capabilities, the application makes data interaction more intuitive, eliminating the need for extensive technical expertise.

The application supports CSV and Excel file uploads, ensuring seamless integration with industry-standard formats. Once uploaded, the system automates data cleaning, exploration, and visualization, allowing users to quickly identify trends and patterns. Automated data cleaning tools detect and correct inconsistencies, ensuring data accuracy and reliability before analysis.

By leveraging natural language processing (NLP), the application revolutionizes the way users engage with their data. Instead of navigating complex query languages or analytical tools, users can ask questions in plain language, making data analysis more accessible to non-technical users. The intuitive interface enables seamless navigation, allowing users to focus on uncovering insights rather than grappling with technical complexities.

The integration of LLM-driven insights ensures that the

generated outputs are not only relevant but also contextually meaningful and easy to interpret. Users can query their datasets to identify trends, correlations, and anomalies, empowering them to make well-informed decisions. This approach reduces the learning curve typically associated with data analytics, making it more accessible to a broader audience.

Prioritizing usability and accessibility, this project fosters an environment where users can confidently interact with their data in real-time. By removing barriers to data analysis, the application encourages exploration, experimentation, and informed decision-making.

By democratizing data analysis, this project bridges the gap between complex datasets and user-friendly insights. It transforms the way users interact with data, enabling them to optimize operations, uncover new opportunities, and foster a data-driven mindset. Ultimately, this application empowers businesses, researchers, and individuals alike to harness the power of AI-driven analytics for smarter decision-making.

## II. LITERATURE SURVEY

The paper by Dr. Sudha SV, Sunil S K, Parthiv Akilesh A S, and Satish G (2024), titled "Democratizing Data Science: Using Language Models for Intuitive Data Insights and Visualizations," focuses on how language models can be employed to make data science more accessible to non-experts. Presented at an IEEE conference, the study explores the potential of leveraging advanced language models to simplify data interaction by enabling users to generate insights and visualizations through natural language queries. The authors highlight how this approach reduces the need for specialized technical skills, allowing a broader audience to engage with data-driven decision-making. Through this work, the paper addresses the growing demand for user-friendly tools in data science and emphasizes the role of AI in democratizing access to complex data analytics and visualizations.[1]

The paper by Liqi Cheng, Dazhen Deng, Xiao Xie, Rihong Qiu, Mingliang Xu, and Yingcai Wu (2021), titled "SNIL: Generating Sports News from Insights with Large Language Models," explores the application of large language models (LLMs) for generating sports news articles based on data-driven insights. Published in the IEEE journal, the study introduces the SNIL framework, which leverages LLMs to transform structured sports data into coherent and engaging news stories. The authors demonstrate how their model can automatically produce human-like narratives by extracting key insights from sports events, enabling fast and efficient news generation. This approach not only enhances the speed of content production but also addresses the challenges of creating accurate, relevant, and contextually appropriate sports reports, significantly advancing the field of automated journalism.[2]

The paper titled "Natural Language Interfaces for Tabular Data Querying and Visualization: A Survey," published in IEEE Transactions on Knowledge and Data Engineering, provides a comprehensive overview of the development and advancements in natural language interfaces (NLIs) designed for querying and visualizing tabular data. The survey discusses

various systems and methodologies that enable users to interact with data using natural language, eliminating the need for complex querying languages or programming skills. It examines the challenges, such as ambiguity in natural language, and the techniques used to address them, including the integration of machine learning models and rule-based approaches. The paper also highlights the evolution of NLIs, the current state of the technology, and its potential to democratize data analysis and visualization, making it more accessible to non-expert users.[3]

The paper titled "Language-Driven Visualization Design: A Study of LLMs in Interactive Data Exploration," presented at IEEE VIS 2024, investigates the role of large language models (LLMs) in facilitating interactive data exploration and visualization design. The authors explore how LLMs can be utilized to enable users to create and customize data visualizations through natural language commands, thus reducing the need for technical expertise in data visualization tools. The study delves into the effectiveness of LLMs in understanding user queries, generating appropriate visualizations, and offering real-time feedback. By integrating LLMs into the data exploration process, the paper highlights the potential of language-driven interfaces to streamline visualization workflows and make data interaction more intuitive and accessible for both experts and non-experts alike. [4]

The paper by Perozzi et al. (2024), titled "Let Your Graph Do the Talking: Encoding Structured Data for LLMs," explores innovative methods for encoding structured data to enhance the capabilities of large language models (LLMs). The authors argue that effectively representing structured data in a way that LLMs can understand is crucial for improving their performance in generating insights and responses. The study introduces various encoding techniques and evaluates their effectiveness in enabling LLMs to leverage the inherent relationships and structures present in the data. By focusing on how structured data can be transformed into formats that LLMs can interpret, the paper aims to bridge the gap between traditional data representation methods and the advanced processing capabilities of LLMs, ultimately enhancing their ability to facilitate data-driven conversations and insights.[5]

The paper by Vertsel and Rumiantsau (2024), titled "Hybrid LLM/Rule-based Approaches to Business Insights Generation from Structured Data," presents a novel framework that combines large language models (LLMs) with rule-based systems to extract valuable business insights from structured data. The authors highlight the strengths and limitations of both approaches, advocating for a hybrid methodology that leverages the precision of rule-based systems for specific tasks alongside the flexibility and natural language processing capabilities of LLMs. This dual approach aims to improve the efficiency and accuracy of generating actionable insights, particularly in complex business scenarios. By addressing the shortcomings of traditional methods, the paper provides a significant contribution to the field of business intelligence, demonstrating how the integration of LLMs can enhance data

analysis and decision-making processes.[6]

Additionally, this chapter explores the latest advancements in natural language processing (NLP) and large language models (LLMs), investigating how these technologies can be harnessed to make data interaction more intuitive and accessible for a broader range of users. Through this exploration, we emphasize the importance of our proposed solution in bridging these identified gaps.

## III. COMPARISON

The AI-powered Data Insights and Visualization application provides an innovative approach to data analysis, allowing users to seamlessly explore, process, and visualize large datasets from any location with an internet connection. Unlike traditional data analytics tools that often require extensive technical expertise and complex configurations, this web-based platform is designed with usability and accessibility in mind. Users can upload datasets in popular formats such as CSV and Excel, making it highly adaptable across various industries and professional settings. This broad compatibility ensures that users can integrate their existing data workflows without the need for additional software or extensive reformatting.

One of the key differentiators of this application is its advanced natural language processing (NLP) capabilities, which allow users to interact with their data using simple, intuitive queries. Rather than relying on complex SQL commands, programming knowledge, or specialized data analytics training, users can simply type out questions in plain language and receive meaningful insights instantly. This functionality removes technical barriers that often prevent non-experts from engaging with data-driven decision-making, thereby democratizing data analysis and expanding its accessibility to a much broader audience.

Beyond its ease of use, the application also provides significant cost and time savings compared to conventional data analysis methods. Traditional approaches often require manual data entry, extensive spreadsheet manipulation, and repetitive reporting efforts, all of which consume valuable time and increase the likelihood of human errors. By automating essential data processing tasks, such as data cleaning, anomaly detection, statistical analysis, and visualization generation, the application ensures that users can focus on extracting insights rather than managing data complexities. This automation not only reduces the time spent on data preparation but also eliminates inefficiencies associated with manual reporting, making the entire workflow significantly more productive.

Additionally, the application's real-time data processing capabilities ensure that users always have access to the most up-to-date, accurate insights. Unlike static reports that quickly become outdated, this system provides continuous access to live data, enabling organizations to make proactive, data-driven decisions rather than relying on retrospective analysis. This real-time accessibility is especially beneficial in fast-paced environments where timely decision-making can have a direct impact on business performance, operational efficiency, and competitive advantage.

Furthermore, remote accessibility allows users to engage with their data from virtually anywhere, breaking down geographical barriers and fostering collaboration across teams and departments. Whether a user is in an office setting, working remotely, or managing operations on the go, they can easily retrieve, analyze, and share data-driven insights, enhancing overall efficiency and teamwork. This flexibility ensures that decision-making is no longer restricted to specific locations or devices, making data insights more readily available whenever and wherever they are needed.

By combining ease of use, automation, real-time analytics, and universal accessibility, this application serves as a powerful tool for transforming raw data into actionable insights. It empowers individuals, businesses, and organizations to fully leverage their data for strategic decision-making, ultimately driving greater efficiency, accuracy, and informed decision-making across various domains. The ability to effortlessly analyze data, generate visualizations, and gain valuable insights without requiring extensive technical expertise makes this application a game-changer in the field of data analytics, offering unparalleled value to users seeking efficient and intelligent data solutions.

## IV. REVIEW OF EXISTING SYSTEM

The rapid expansion of data-driven decision-making has led to the emergence of numerous data analysis and visualization tools, each catering to specific user needs. Among the most widely used solutions are traditional spreadsheet applications such as Microsoft Excel and Google Sheets, which have long been relied upon for data entry, calculations, and visualization. These platforms offer robust functionalities such as pivot tables, formula-based computations, and basic charting tools, making them indispensable for many businesses and professionals. However, they come with inherent limitations—creating complex formulas, handling large datasets, and performing advanced analytics require significant technical proficiency. Additionally, manual data cleaning and manipulation can be tedious and error-prone, often leading to inefficiencies and inaccuracies in data-driven decision-making.

To address the limitations of spreadsheets, Business Intelligence (BI) tools such as Tableau, Power BI, and QlikView have gained popularity. These tools offer advanced visualization capabilities, interactive dashboards, and in-depth analytics, making them powerful solutions for data exploration and reporting. However, their steep learning curves, high subscription costs, and the need for extensive setup and configuration can pose significant barriers to adoption, particularly for small businesses, non-technical users, and budget-conscious organizations. Additionally, maintaining and updating BI tools requires ongoing investment, both in terms of financial resources and technical expertise, making them less accessible to users who require quick and cost-effective data insights.

Some organizations opt for custom-built data analysis software tailored to their specific business needs. While these solutions provide highly personalized features and functionalities, they come with several drawbacks, including lengthy development timelines, significant financial investment, and the requirement for skilled professionals to build and maintain the system. The high cost and complexity associated with custom solutions make

them infeasible for many businesses, particularly those looking for scalable and flexible data analytics solutions.

In response to the need for more accessible data visualization tools, several cloud-based platforms, such as Google Data Studio and Datawrapper, have emerged. These platforms allow users to connect to various data sources and create visualizations quickly and efficiently without requiring specialized software installations. However, they often lack comprehensive data analysis capabilities, focusing primarily on charting and reporting rather than in-depth data exploration. Moreover, limited data cleaning functionalities make it challenging to ensure accuracy and reliability, requiring users to preprocess their data manually before utilizing these tools.

Another recent advancement in data analysis tools is the integration of Natural Language Processing (NLP) technologies, which enable users to interact with their data using simple, conversational queries. While promising, many NLP-driven solutions remain in their early stages of development, facing challenges related to query accuracy, limited contextual understanding, and compatibility with diverse data formats. Additionally, many NLP-based tools struggle to generate meaningful insights beyond simple summaries, making them insufficient for complex data analysis and visualization needs.

Despite the variety of existing data analytics tools, each comes with trade-offs that limit accessibility, usability, and efficiency. Many tools are either too complex for non-technical users, too expensive for small businesses, or lack key features such as automated data cleaning, real-time processing, and natural language interaction. As a result, there remains a significant gap in the market for a comprehensive, user-friendly data analysis and visualization solution that combines the intuitive nature of NLP, the analytical power of BI tools, and the flexibility of cloud-based platforms.

The proposed Data Insights and Visualization application aims to fill this gap by offering a seamless, AI-driven experience that allows users to upload, clean, explore, and visualize data effortlessly. By leveraging Large Language Models (LLMs), the application simplifies complex data interactions, enabling users to generate insights through conversational queries without requiring specialized skills. This approach ensures that both technical and non-technical users can access, analyze, and act on their data efficiently, ultimately enhancing productivity, decision-making, and overall business intelligence.

## V .METHODOLOGY

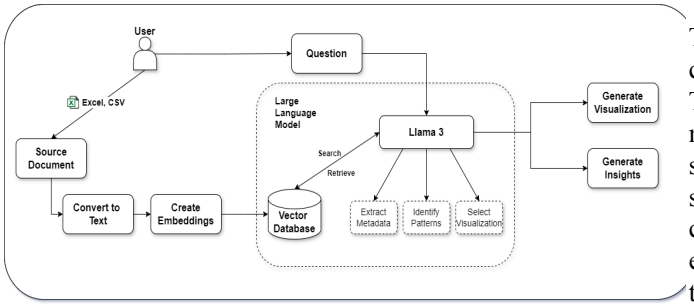


*fig 5.1 Architectural Framework*

This diagram represents a system where a user queries structured data (e.g., Excel or CSV files) to gain insights and visualizations using a Large Language Model (Llama 3). The source documents are first converted to text, and embeddings (numeric representations) are created and stored in a vector database. When a user asks a question, the LLM searches the database to extract relevant metadata, identify patterns, and select suitable visualizations. The system then generates insights and visualizations based on the retrieved information, providing answers in a meaningful and graphical form.

- Data Collection

Flexible Data Formats:

The platform is designed to accommodate a wide range of input data formats, making it versatile and user-friendly for various types of users. It can accept files in formats such as Excel, CSV, or other common structured data types. This flexibility ensures that users from different industries or domains can seamlessly upload and process their data without needing to convert files into a specific format. For instance, an Excel sheet with multiple tabs or a CSV with thousands of rows can be ingested into the system, preserving the structure and data integrity. This broad support reduces friction in data handling and allows users to focus on analyzing the data rather than on preprocessing it.

- Data Analysis Using LLMs

Advanced Feature Extraction:

Leveraging Large Language Models (LLMs) allows for the extraction of complex and high-level features from the uploaded data. Beyond simple statistical measures, the LLM can detect advanced patterns such as time series trends, cyclical behaviors, and hierarchical relationships within the data. For instance, in a time series dataset, the LLM could automatically detect seasonality, trends, and recurring patterns, offering insights that traditional models might miss. If the data contains hierarchical structures, like nested categories, the LLM can identify and maintain these relationships during analysis. This advanced extraction capability enriches the analysis process and makes it more sophisticated, enabling deeper and more meaningful insights.

- Contextual Understanding:

The LLM is equipped with the ability to interpret the data contextually, using domain-specific knowledge when necessary. This feature is crucial for generating insights that are not only mathematically correct but also relevant and actionable within a specific domain. For example, if the data pertains to the financial sector, the LLM can apply its understanding of financial concepts (such as market cycles, economic indicators, etc.) to enhance its analysis. This contextual awareness allows the LLM to provide insights that are more accurate and tailored to the user's needs, ensuring that the data's meaning is preserved and interpreted correctly.

- Automated Chart Creation

Dynamic Chart Selection:

The system is designed to intelligently select the most suitable type of visualization based on the characteristics of the data and the user's specific query, rather than relying on static, predefined rules. This dynamic approach means that the system can evaluate the nature of the data—whether it's time-series data, categorical data, or numerical distributions—and choose an appropriate chart type, such as a line chart, bar chart, or heatmap, respectively. For instance, if the user's data contains two variables with a strong correlation, the system may suggest a scatter plot, while hierarchical data might lead to the creation of a tree map. This flexibility enhances the user experience, ensuring that visualizations are informative and contextually appropriate.

- Data-Driven Recommendations:

In addition to selecting charts based on the current data and query, the system can also provide proactive recommendations for additional visualizations that the user may not have initially considered. For example, if the user is analyzing sales data and generates a time-series graph, the system could suggest other visualizations like a pie chart to display the distribution of sales across regions or a histogram to show sales performance by category. These recommendations can expand the user's exploration of the data and encourage a deeper understanding of potential insights that might otherwise be overlooked.

- User-Friendly Visualization

Export and Sharing Options:

To enhance usability, the platform allows users to export their visualizations in various file formats, such as PNG, PDF, or even vector formats like SVG for high-resolution needs. These export options ensure that users can share their findings in a format that best suits their presentation or reporting needs, whether for business meetings, academic papers, or internal reports. Additionally, the platform could offer direct sharing features, enabling users to send visualizations via email or to integrate them into collaborative tools like Slack or Microsoft Teams. This ease of export and sharing ensures that insights generated on the platform can be disseminated quickly and effectively, supporting better decision-making.

## VI. RESULT

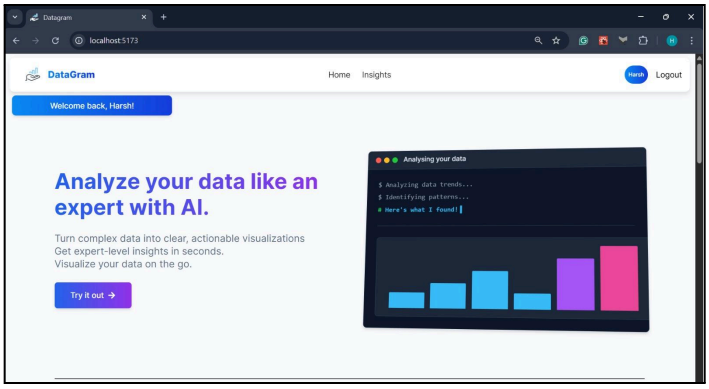


*fig 6.1.1 Home Screen*

This is the homepage of **DataGram**, a smart AI-powered platform that transforms complex datasets into clear, actionable visualizations—making expert-level analysis simple and fast.

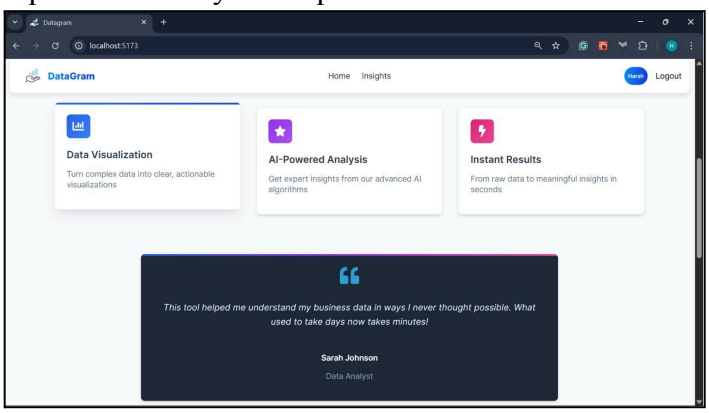


*fig 6.1.2 Home Screen (Scrolled view)*

This section highlights DataGram's core features—data visualization, AI-powered analysis, and instant results—showcasing how users can turn raw data into meaningful insights within seconds.

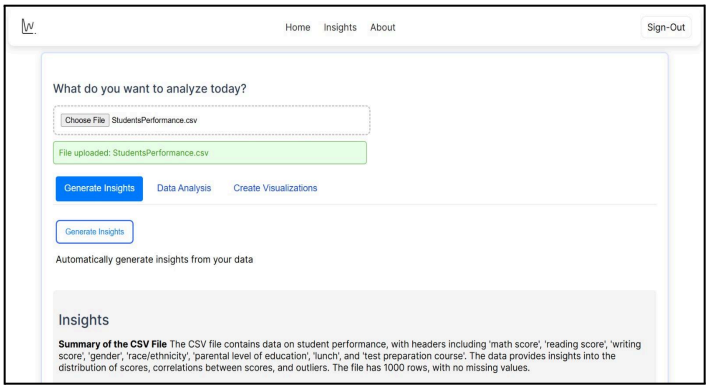


*fig 6.1.3 Insights Generation*

This screen demonstrates DataGram's intelligent insight generation feature. After uploading a dataset (StudentsPerformance.csv), users can instantly generate meaningful insights, such as column

summaries, data distribution, correlations, and outlier detection—all without writing a single line of code.

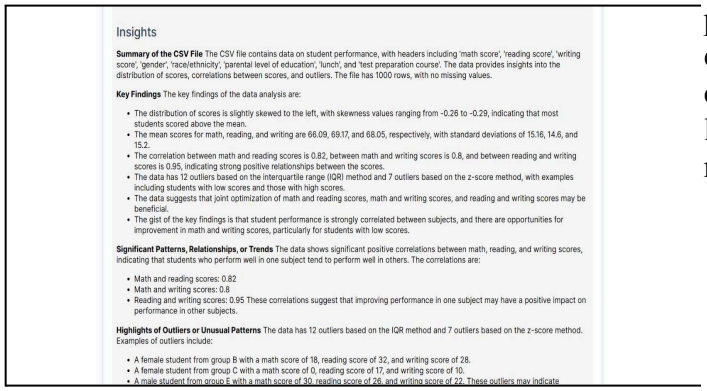


*fig 6.1.4 Insights Display*

This screen showcases DataGram's automated insight explanation module. After analyzing the uploaded dataset, it generates a detailed summary highlighting:

- Key Findings like skewness in score distributions, subject-wise mean and standard deviation, and strong correlations among subjects (e.g., math and writing: 0.8).
- Significant Patterns & Trends such as students performing well in one subject tending to excel in others.
- Outlier Detection identifying specific students with unusual score combinations.

This feature helps users quickly grasp the core insights of their data—turning raw numbers into understandable narratives.

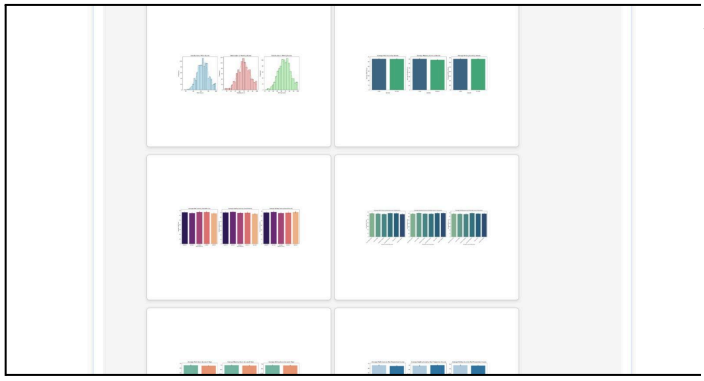

*fig 6.1.5 Insights Visualizations*

The auto-generated visualizations from DataGram provide key insights into student performance. Distribution plots reveal the spread of scores in reading, writing, and math, helping identify overall performance trends. Gender-wise comparisons highlight that average scores vary slightly between male and female students. The impact of parental education is also evident, with higher education levels generally correlating with better student outcomes. Further, the ethnicity-wise breakdown showcases performance trends across different racial groups. The effect of test preparation is clear, as students who completed a prep course tend to score higher. Lastly, lunch type—standard vs. free/reduced—also shows a noticeable impact on student performance.

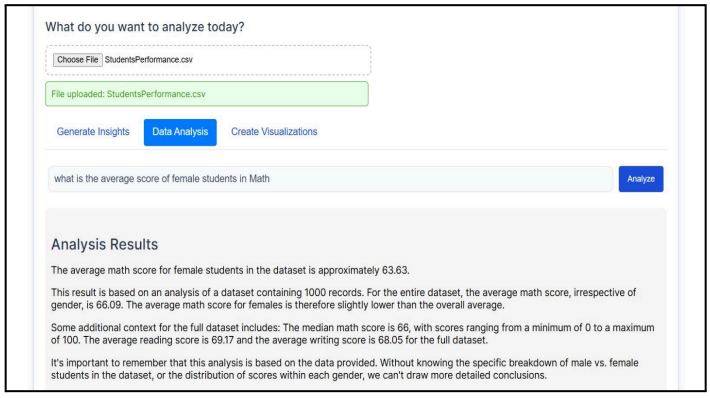


*fig 6.1.6 Chat to Analysis & the result*

The analysis reveals that the average math score for female students in the dataset is approximately 63.63. This value is slightly below the overall average math score of 66.09 across all students in the dataset of 1,000 records. The median math score is 66, with scores ranging from 0 to 100. Additionally, the average reading score is 69.17, and the average writing score is 68.05 for the full dataset. It's important to note that this analysis is based solely on the provided data, and more detailed conclusions would require a deeper breakdown of scores by gender and other variables.

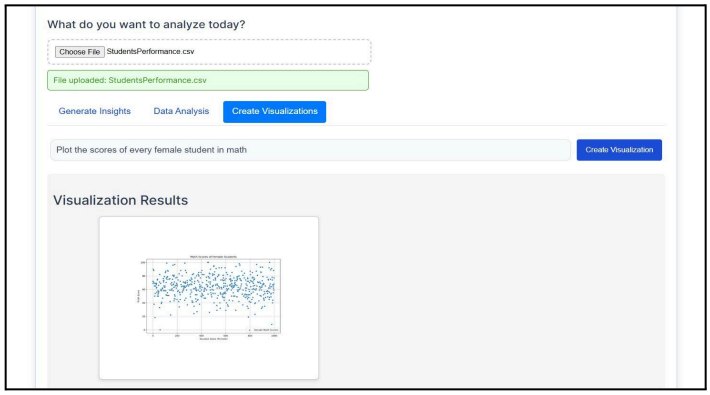


*fig 6.1.7 Chat to Visualization & the result*

The scatter plot shows math scores of female students, with most scoring between 50 and 80. This suggests generally consistent performance among them.

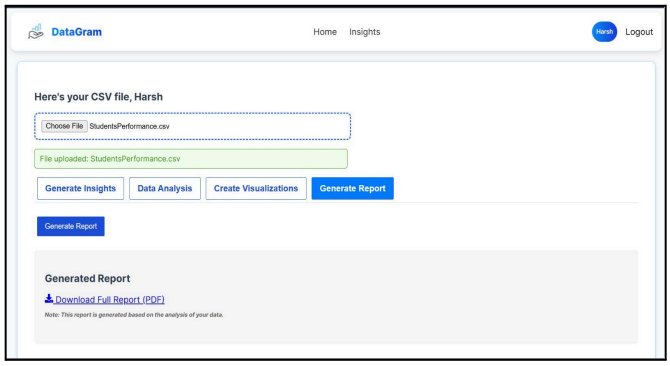


*fig 6.1.8 Generate Report*

The platform allows users to upload CSV files for automated analysis. After uploading a file (e.g., *StudentsPerformance.csv*), users can generate insights, perform data analysis, create visualizations, and generate a downloadable PDF report based on the uploaded data.

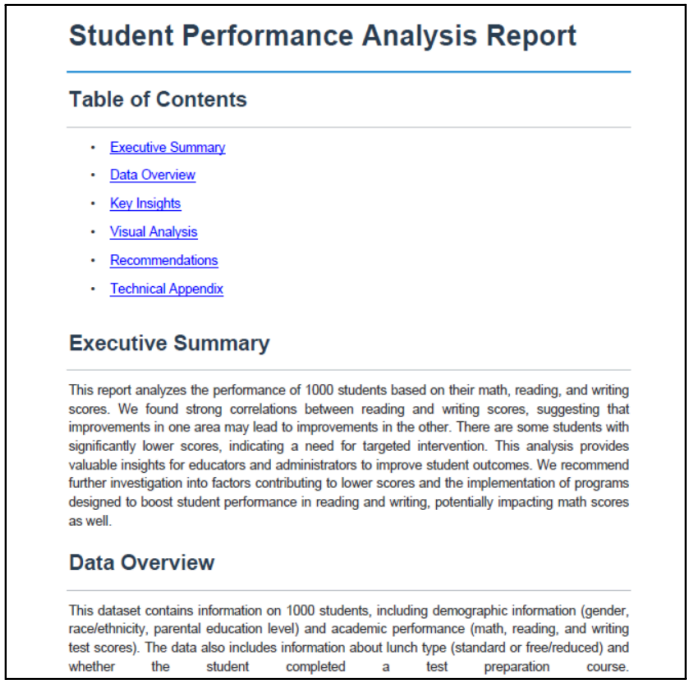


*fig 6.1.9 Report*

After uploading the CSV file, the platform generated a detailed *Student Performance Analysis Report*. The report includes an executive summary, data overview, key insights, visual analysis, recommendations, and a technical appendix. It provides a comprehensive analysis of student performance, helping educators identify trends, correlations, and areas for targeted improvement.

## VIII. 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.

[8]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.

[9]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

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

[11]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.

[12]"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