CAPSTONE PROJECT

ANALYZING IMDB MOVIE DATASET FOR INSIGHTS AND TRENDS

Presented By:
JAGAN MOHAN AITHEPALLI - Annamacharya institute of Technology & Sciences –Artificial Intelligence and Data science

edunet foundation

OUTLINE

- **Problem Statement** (Should not include solution)
- **Proposed System/Solution**
- **System Development Approach** (Technology Used)
- **Result**
- **Conclusion**
- **Future Scope**
- **References**

PROBLEM STATEMENT

With the vast amount of data available on IMDb, there exists an opportunity to gain valuable insights into the world of movies. The problem at hand is to conduct a comprehensive analysis of the IMDb movie dataset to uncover trends, patterns, and correlations within the data. This analysis aims to provide valuable insights to various stakeholders in the movie industry, including filmmakers, producers, distributors, and investors.

PROPOSED SOLUTION

- The proposed system aims to address the challenge of analyzing the IMDb movie dataset to extract valuable insights and trends. It involves leveraging data analytics and potentially machine learning techniques to uncover patterns and correlations within the dataset. The solution will consist of the following components:
- Data Collection:
- Gather IMDb movie dataset, including information such as movie title, genre, release year, ratings, and metascore.
- Augment dataset with additional relevant data sources, such as budget, revenue, director, and cast information.
- Ensure data integrity and quality through thorough validation and cleansing processes.
- Data Preprocessing:
- Cleanse and preprocess the collected data to handle missing values, outliers, and inconsistencies.
- Perform feature engineering to extract relevant features from the dataset, such as genre trends over time, directorial influence, and budget-revenue relationships.

CONTINUED

- Exploratory Data Analysis (EDA):
- Conduct exploratory analysis to visualize distributions, correlations, and trends within the dataset.
- Explore relationships between different variables, such as budget and revenue, ratings and genre, and directorial impact on movie success.
- Deployment:
- Develop a user-friendly interface or application that provides interactive visualizations and insights derived from the IMDb dataset.
- Deploy the solution on a scalable and reliable platform, ensuring accessibility and usability for stakeholders in the movie industry.
- Evaluation:
- Assess the effectiveness of the analysis and insights provided by the system through stakeholder feedback and validation against known industry trends.
- Continuously monitor and update the system to incorporate new data and improve prediction accuracy and relevance.

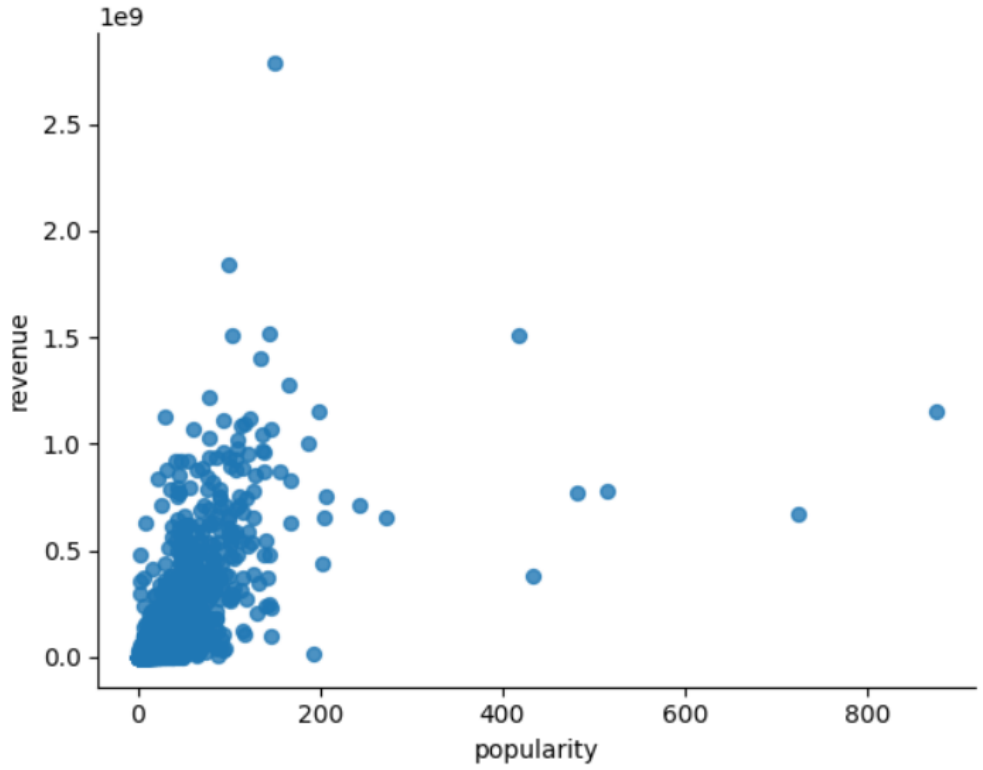
SYSTEM APPROACH

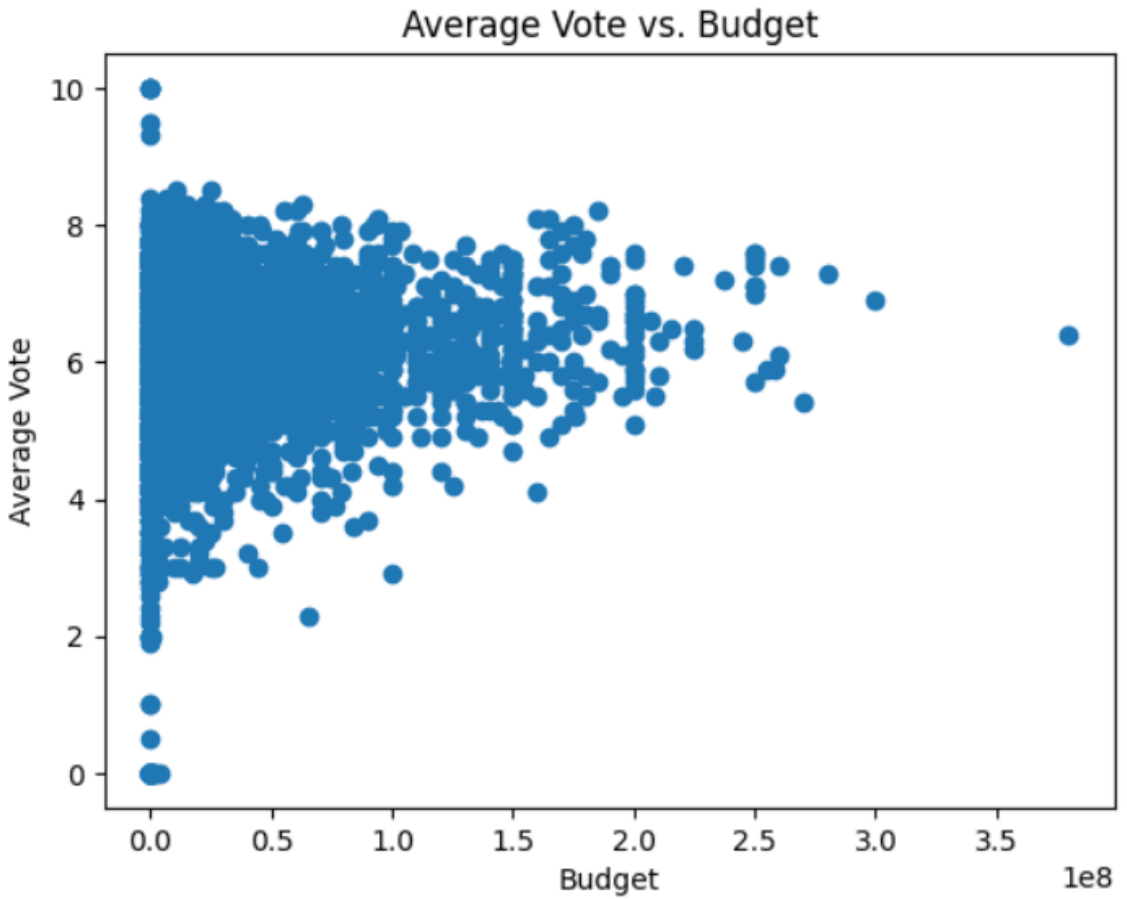
- System Requirements: The system for IMDb dataset analysis will require computing resources capable of handling data processing, modeling, and visualization tasks. Additionally, access to the IMDb dataset or relevant data sources is essential. The system should be able to handle large datasets efficiently to ensure timely analysis and insights generation. Furthermore, a user-friendly interface or application may be developed to facilitate interaction with the analysis results.
- Libraries Required to Build the Model:
- Pandas: Pandas is a powerful library in Python used for data manipulation and analysis. It will be utilized for loading, cleaning, and preprocessing the IMDb dataset.
- Matplotlib and Seaborn: These libraries are essential for data visualization and will be used to create informative plots and graphs to visualize trends and patterns within the dataset.
- NumPy: NumPy is a fundamental package for scientific computing with Python, providing support for large, multi-dimensional arrays and matrices. It will be used for numerical operations and computations within the analysis.

RESULT

Our exploration of trends within the IMDb movie dataset revealed several noteworthy findings. Firstly, drama and comedy emerged as the most popular movie genres overall, reflecting audience preferences for emotionally engaging narratives and light-hearted entertainment. Secondly, while the average vote showed a gradual decline after the mid-1970s, there was a significant increase in the average voter count over the years, indicating a growing engagement with IMDb as a platform for rating and reviewing movies. Lastly, top directors such as Spielberg, Clint Eastwood, and Ridley Scott stood out for directing a higher number of movies compared to their peers, suggesting their enduring influence and prolificacy in the film industry..

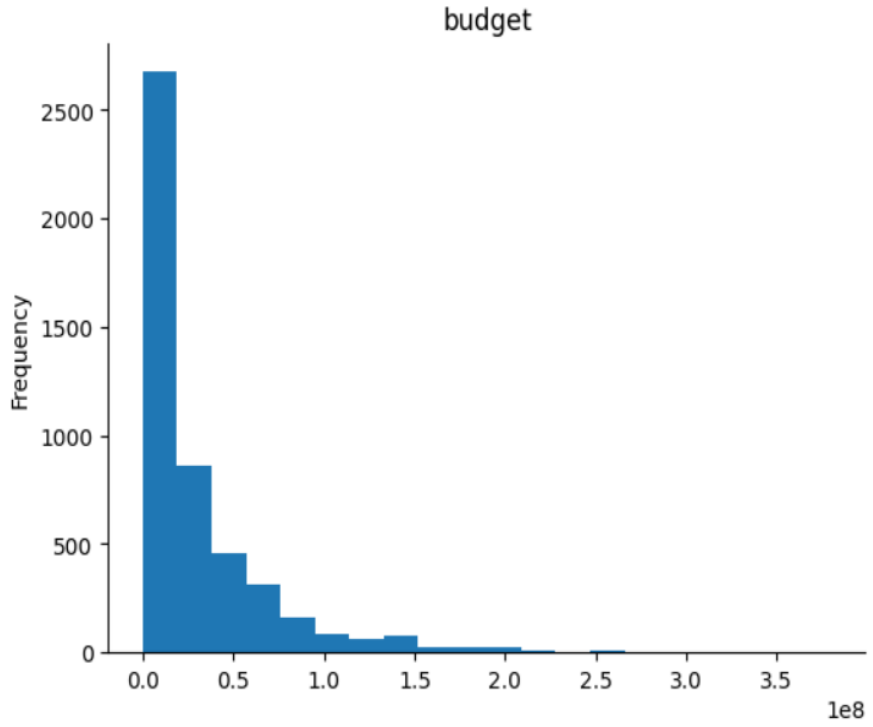
CONTINUED



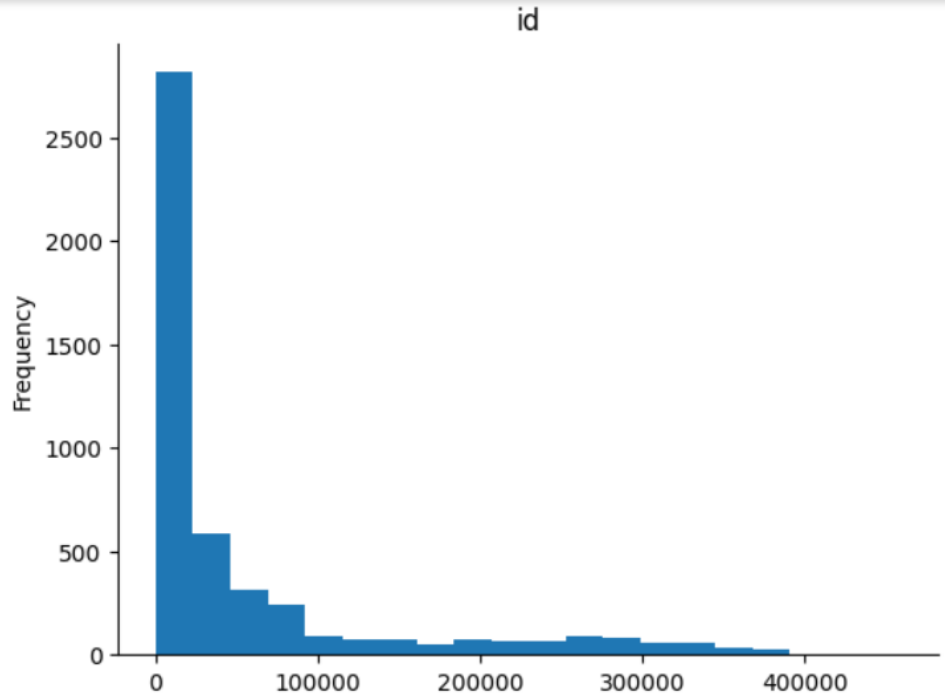


CONTINUED

[14]

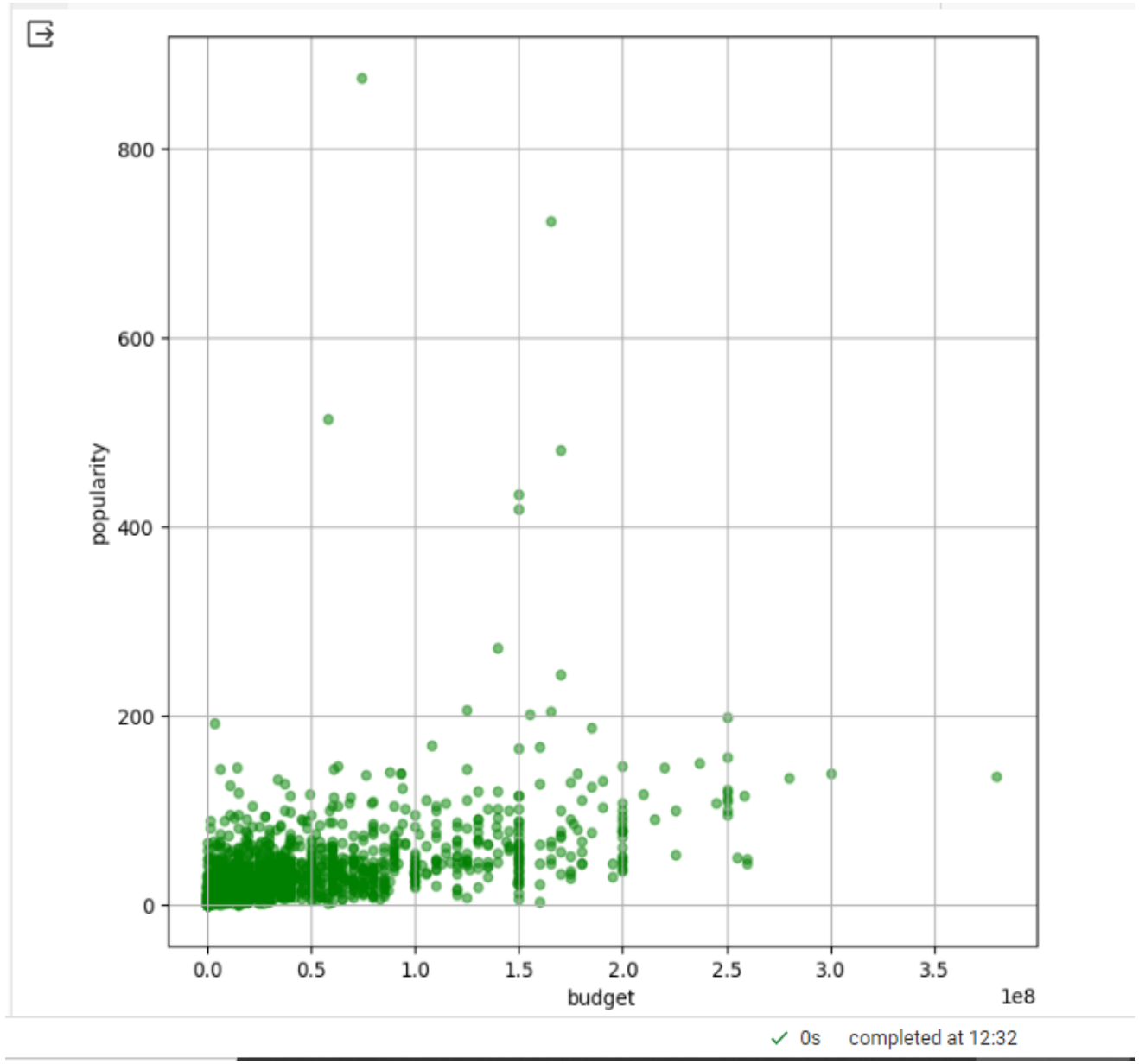


[15]



CONTINUED

CONTINUED



CONCLUSION

- In conclusion, our analysis aimed to identify features correlated with movie revenue in the IMDb dataset. We found a significant positive correlation between movie budget and revenue, suggesting that higher budget films tend to yield greater financial returns. However, the reliability of this conclusion was hindered by data limitations, including a substantial number of missing values and a reduction in sample size after cleaning. Despite these challenges, the correlation between popularity and revenue was evident, reinforcing the intuitive understanding that more popular movies often generate higher revenue. Furthermore, our analysis highlighted specific directors such as Peter Jackson, Steven Spielberg, and Michael Bay, whose films consistently garnered higher revenue compared to others. While this suggests a relationship between directorial reputation and revenue, further investigation is needed to establish causality.

CONCLUSION

- Additionally, trends in revenue over the years revealed fluctuations, with a notable spike in the mid-1960s to 1980s followed by relative stability. However, the credibility of this observation was compromised by the significant number of missing values in the data from the 1960s to 1980s. Despite these limitations, our analysis provides valuable insights into the factors influencing movie revenue, offering filmmakers, producers, and investors a basis for decision-making and strategy formulation within the competitive movie industry landscape. Moving forward, addressing data quality issues and conducting more extensive analyses could enhance the robustness of our findings and contribute to a deeper understanding of the complex dynamics driving movie success.

FUTURE SCOPE

- Building upon these findings, future research could delve deeper into the evolving landscape of movie genres and audience preferences, exploring emerging trends and potential shifts in consumer behavior. Additionally, further investigation into the factors driving the increase in average voter count on IMDb could provide insights into the platform's growing influence and its role in shaping movie perceptions. Moreover, conducting a comparative analysis of the directing styles and thematic preferences of prolific directors like Spielberg, Eastwood, and Scott could offer valuable insights into their creative processes and the factors contributing to their continued success. By expanding upon these trends and exploring their implications, future studies can contribute to a more comprehensive understanding of the dynamics shaping the contemporary film industry.

REFERENCES

1)Please note that the 'explode' function used to separate the movie genres from the genre columns was taken from the following stack overflow question
https://stackoverflow.com/questions/12680754/split-explode-pandas-dataframe-string-entry-to-separate-rows
2) I also drew inspiraton from the following Kaggle analysis. It was interesting to note that the original dataset contains far more features and the limitless possiblities of data wrangling that can be done.
https://www.kaggle.com/aninda123/imdb-movie-analysis

COURSE CERTIFICATE 1

In recognition of the commitment to achieve professional excellence

Journey to Cloud: Envisioning Your Solution
IBM SkillsBuild

JAGAN MOHAN AITHEPALLI

Has successfully satisfied the requirements for:

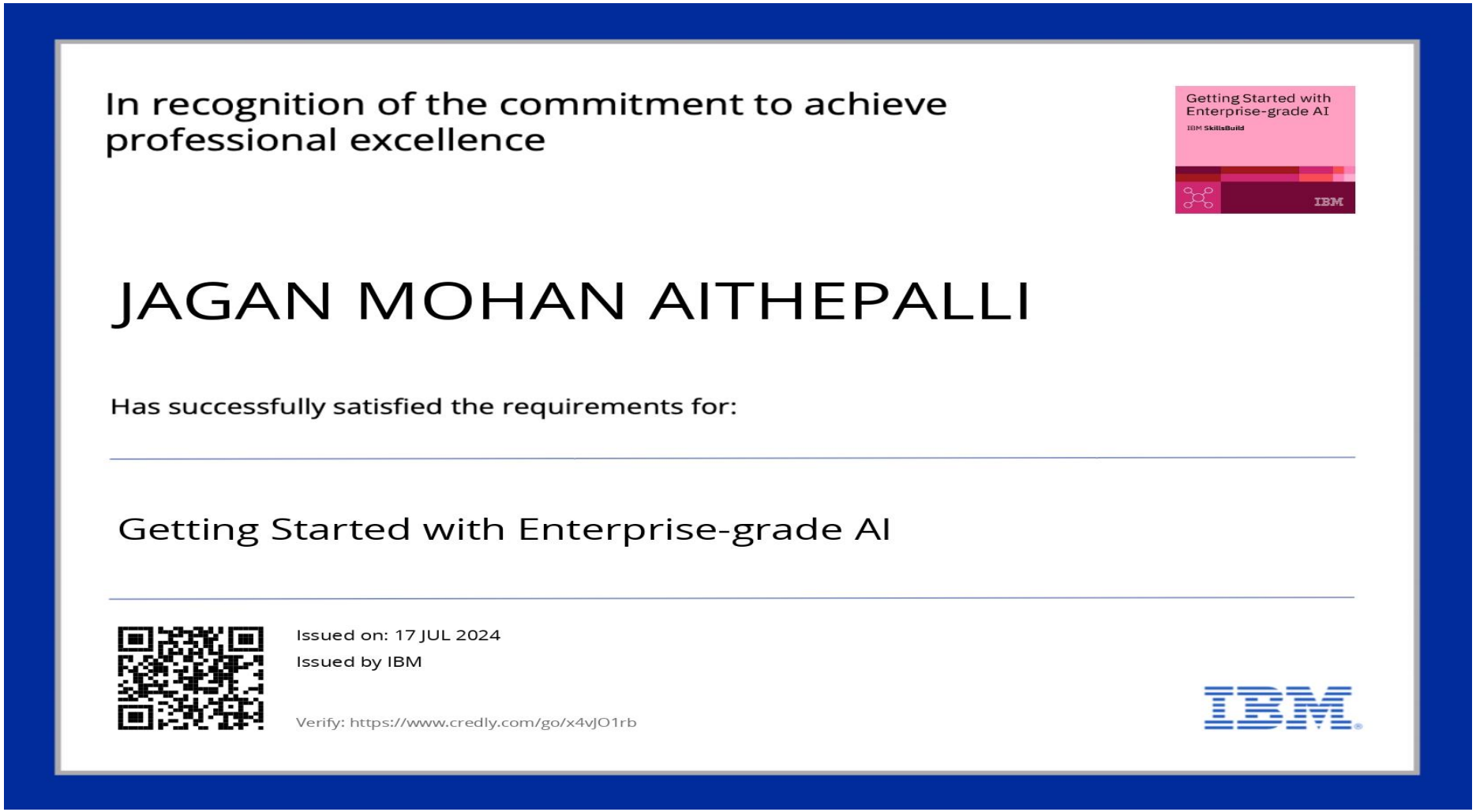
Journey to Cloud: Envisioning Your Solution

Issued on: 17 JUL 2024
Issued by IBM

Verify: https://www.credly.com/go/C7tD3gry

IBM

COURSE CERTIFICATE 2



edunet foundation

THANK YOU

edunet foundation