Data Engineering Graduation Project Report

Project Title: Movie Review Sentiment Analysis and Improvement

Initiative: Digital Egypt Pioneers Initiative - DEPI, AI & Data Science Track

Team Members and Roles:

- **Haneen Alaa:** Manages data acquisition processes and performs big data analysis to support comprehensive data-driven decision-making.
- **Mirna Sherif:** Responsible for designing and managing the ETL processes and overseeing the data warehousing architecture.
- **Alaa Atef:** Focused on writing complex SQL queries for data validation and analysis.
- **Gehad Emad:** Develops interactive data visualizations using Power BI and cleans the data.
- **Salma Essam:** Handles the deployment of machine learning models and manages the integration of ML solutions into the production environment.

Project Duration: 4 Weeks

1. Introduction

The purpose of this project is to analyze movie statistics and viewer preferences, providing valuable insights that help producers and streaming platforms tailor future releases. By understanding trends in genres, runtimes, and viewer feedback, producers can make informed decisions to align their content with audience expectations and trends, ensuring better engagement and success.

2. Project Need

The film industry generates vast amounts of data, with numerous movies and millions of reviews each year. Analyzing this data manually is challenging due to the volume and complexity of viewer preferences, which are highly diverse and region-specific. This project aims to uncover insights from a massive dataset of movies and reviews, addressing the challenges of data complexity and enabling producers to make data-driven decisions.

Key Challenges Addressed:

- **Massive Dataset of Movies & Reviews:** Handling and analyzing a large volume of titles and reviews.
- **Complex Viewer Preferences:** Identifying genre, runtime, and production elements that resonate with different audiences.
- **Untapped Data Insights:** Structuring the data to uncover hidden insights about viewer preferences and genre success.

Solution:

- **Data-Driven Insights:** By analyzing historical data, this project provides actionable insights that help improve the quality and appeal of future films.

- **Comprehensive Dataset Evaluation:** The analysis draws from large datasets, including IMDb, to capture trends and accurately interpret audience reactions.

3. Data Flow and Components

- **Data Acquisition:** Data was obtained through web scraping and other sources.
- **Data Cleaning and Preprocessing:** Cleaning various data tables, such as alternative movie titles (akas), individual names and their professions (name_basics), and detailed crew information (title_crew).
- **Data Integration:** The cleaned data was used for further analysis, ensuring consistent and accurate insights.

4. Data Warehouse and ETL Processes

The data warehouse serves as the backbone of the project, enabling efficient data extraction, transformation, and loading (ETL). The key steps include:

- **Designing and Managing ETL Processes:** Handling data integration from multiple sources and transforming it for consistency.
- **Creating Dimensions:** Establishing dimension tables for movies, genres, and regions to facilitate more detailed analysis.

5. SQL Queries and Analysis

Complex SQL queries were used to validate and analyze the data. Key insights were extracted, such as the most liked genres, successful runtime combinations, and trends across regions. This analysis helped identify which genres and combinations were most popular.

6. Big Data Analysis and Insights

The project also involved big data analysis to derive insights from large datasets, such as:

- **Identifying Popular Genres in Regions:** Understanding which genres perform well in specific regions.
- **Genre Trends and Viewer Preferences:** Analyzing trends in movie genres and how they relate to viewer satisfaction.

7. Data Visualization

- **Interactive Dashboards:** Power BI was used to create interactive dashboards, providing stakeholders with clear insights into movie trends and viewer preferences.
- **Key Metrics:** Metrics such as genre popularity, average viewer rating, and preferred runtime were visualized, enabling an easy interpretation of complex data.

8. Deployment and Machine Learning Integration

- **Deployment of ML Models:** Machine learning models were deployed to classify movie reviews and predict viewer sentiment. This step ensured that the insights were actionable in a production environment, directly contributing to decision-making processes for content creation.

9. Conclusion

The Movie Review Sentiment Analysis project successfully demonstrated how data engineering and data science can be leveraged to understand and improve movie production strategies. By addressing the challenges of data volume and complexity, the project delivered actionable insights into viewer preferences and industry trends.

Future Work:

- **Expansion of Dataset Sources:** Incorporate data from additional platforms for more comprehensive insights.
- **Further ML Model Improvement:** Improve the accuracy of the sentiment analysis models by using more advanced algorithms or expanding training data.

Acknowledgment: We thank the Digital Egypt Pioneers Initiative (DEPI) for their support and guidance throughout the project.