

**ITC 6460 Cloud Analytics**

**Final Project Report – Video Games Market Analysis**

**Group: Husky 2**

By

**- Katherine LaConte**

**- Likhitha Varakala**

**- Manmitha Pantangi**

**- Shivam Sinha**

**Business Questions:**

If you’re going to create a new video game, what type should you create to be successful?

Which platform is most popular across the globe?

Which genre is currently most popular across the globe?

Which publishers are most successful in North America?

Which games in the past 10 years had a perfect 10 critic rating?

**Data Overview:**

The dataset for this analysis encompasses two comprehensive files: game\_sales\_data.csv and vgsales.csv. These datasets provide an extensive overview of the video game industry, cataloging various aspects such as game titles, platforms, release years, genre classifications, and sales figures across different regions including North America, Europe, Japan, and other territories. This rich compilation of data offers a unique lens through which to view the dynamics of video game sales globally, capturing the trends, preferences, and shifts in the market over time. It serves as a critical foundation for the analytical tasks ahead, enabling a multifaceted exploration of what drives success in the video game industry.

**Cloud Technology:**

AWS Glue played a critical role in the data preparation phase. As a fully managed extract, transform, and load (ETL) service, AWS Glue automated the time-consuming tasks of data cataloging, cleaning, enrichment, and transformation. It turned raw data from S3 into a structured format that was ready for analysis. This automation not only streamlined the data preparation process but also ensured consistency and reliability in the data being analyzed.

Amazon Athena is a serverless, interactive query service provided by Amazon Web Services (AWS). It allows users to analyze data stored in Amazon S3 using standard SQL queries without the need for managing infrastructure. Athena supports various data formats, including JSON, Parquet, ORC, and more. Users can easily create tables, define schema on their data, and execute SQL queries through the Athena console or API. With its pay-as-you-go pricing model, users only pay for the data scanned by their queries, making it cost-effective. Athena is suitable for ad-hoc analysis, data exploration, and extracting valuable insights from large datasets stored in S3.

**Amazon SageMaker** was leveraged for its powerful machine learning capabilities. It enabled the building, training, and deployment of a regression model aimed at predicting global video game sales. SageMaker's comprehensive and user-friendly environment accelerated the development of the model by providing access to high-performance computing resources, pre-built algorithms, and the flexibility to experiment with different model configurations. This facilitated the creation of a predictive model that could accurately forecast sales trends based on historical data and other influencing factors.

**AWS QuickSight** was the final piece in the data analysis pipeline, bringing the insights and findings to life through interactive dashboards and visualizations. QuickSight's ability to connect directly to data stored in AWS services, such as S3 and AWS Glue, allowed for real-time analysis and visualization. The dashboards created provided a dynamic and intuitive interface for exploring the data, enabling stakeholders to quickly understand the analysis outcomes and make informed decisions. By leveraging QuickSight's advanced visualization features, we were able to highlight key trends, patterns, and anomalies in the video game sales data, making the insights accessible to a broad audience.

**Benefits:**

**Scalability:**

AWS Athena and SageMaker are designed to handle large-scale datasets. As your data grows, these services can scale horizontally, accommodating increased computational demands without requiring manual adjustments or infrastructure provisioning.

**Impact:** Enables seamless handling of expanding datasets and ensures consistent performance as data volumes increase.

**Serverless Athena Queries:**

Athena allows users to query data in Amazon S3 without the need for managing servers or infrastructure. It follows a serverless model, enabling users to focus on querying and analyzing data without the overhead of infrastructure management.

**Impact**: Facilitates agile and ad-hoc querying, providing quick insights without the need for provisioning and maintaining dedicated resources.

**Integration with S3:** Athena seamlessly integrates with Amazon S3, a cost-effective and scalable storage solution. This integration simplifies data storage and retrieval processes, allowing for efficient data access directly from S3.

**Impact:** Provides a cohesive ecosystem where data can be stored, queried, and processed efficiently, promoting ease of use and flexibility.

**Diverse Libraries and Frameworks (SageMaker):**

SageMaker supports various machine learning libraries and frameworks, including TensorFlow and Scikit-Learn. This diversity allows data scientists and developers to choose the most suitable tools for their specific modeling requirements.

**Impact**: Enhances flexibility in model development, enabling the use of popular machine learning libraries and frameworks for diverse tasks.

Drawbacks and Challenges:

**Learning Curve:**

Adoption of AWS services may involve a learning curve, especially for users new to cloud computing. Acquiring the necessary skills and understanding of AWS services can take time.

**Impact:** Initial challenges in understanding the services may slow down the onboarding process and necessitate additional training.

**Cost Management:**

While AWS services offer a pay-as-you-go pricing model, it's essential to monitor and manage costs effectively. Usage of resources beyond optimal levels can lead to unexpected expenses.

**Impact:** Inadequate cost management may result in budget overruns, making it crucial to implement cost-control measures and monitoring mechanisms.

**Data Security and Compliance:**

Ensuring the security and compliance of sensitive data, such as video game sales data, is a significant concern. Proper configuration and adherence to security best practices are necessary.

**Impact:** Inadequate security measures may lead to data breaches or non-compliance issues, posing reputational and legal risks.

**Model Interpretability:**

Depending on the complexity of the predictive model built with SageMaker, interpreting the model's results and explaining its predictions can be challenging.

**Impact:** Lack of interpretability may limit stakeholders' understanding of the model's decision-making process, potentially impacting trust and acceptance of model outcomes.