**Big Data Case Study**

**Analyzing Global Video Games Sales Data**

**Introduction**

With the explosion of digital data, industries are turning to Big Data technologies to extract valuable insights from large and complex datasets.  
This case study demonstrates how Hadoop ecosystem tools — HDFS, Sqoop, and Hive — can be used for data ingestion, storage, processing, and analysis.  
We analyze a real-world dataset of video game sales, exploring trends in platforms, publishers, and regional performance.

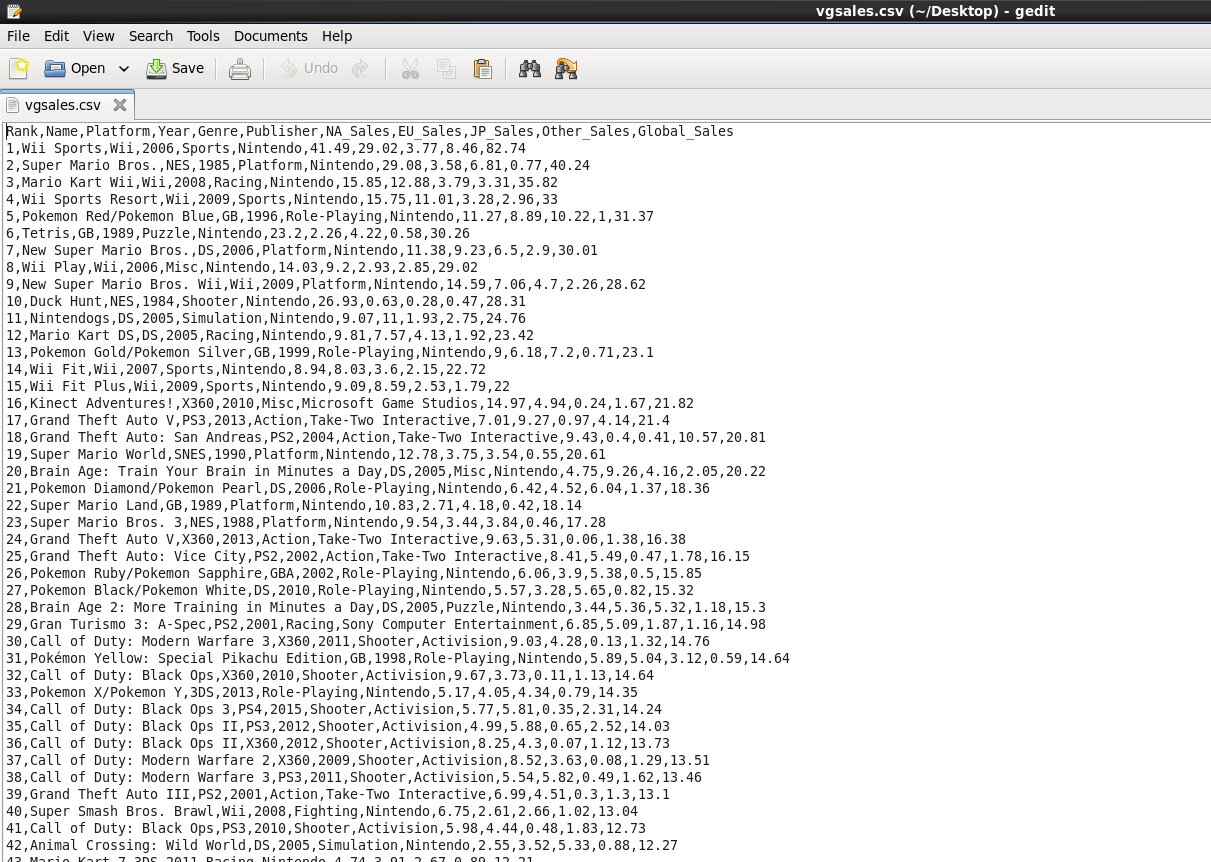
**Details of Dataset**

**Dataset name:** vgsales.csv  
**Source:** Kaggle – Video Game Sales Dataset  
**Description:** Contains sales information of video games across multiple regions.

| **Column Name** | **Description** |
| --- | --- |
| Rank | Rank of the game based on global sales |
| Name | Name of the video game |
| Platform | Platform on which the game was released (e.g., PS4, Xbox, PC) |
| Year | Year of release |
| Genre | Type of game (e.g., Action, Sports, RPG) |
| Publisher | Company that published the game |
| NA\_Sales (million) | Sales in North America |
| EU\_Sales (million) | Sales in Europe |
| JP\_Sales (million) | Sales in Japan |
| Other\_Sales (million) | Sales in other regions |
| Global\_Sales (million) | Total worldwide sales |

**Total records:** ~16,500  
**File format:** CSV (Comma-separated values)

A part of the dataset is:



**Project Scope**

The objective of this project is to:

* Ingest and process the dataset using Hadoop ecosystem components.
* Perform analysis using **Hive** (Big Data warehouse).
* Generate visual insights into sales trends and publisher performance.

**Goals**

1. Import dataset into **HDFS** file system.
2. Store and query data using **Hive**.
3. Perform exploratory and comparative data analysis.
4. Visualize results using analytical tools like **Matplotlib**, **Seaborn etc.**

**Tools and Working Environment**

| **Tool** | **Purpose** |
| --- | --- |
| **HDFS** | Distributed storage for large-scale data |
| **Hive** | Data warehouse for querying using SQL-like syntax |
| **Hadoop** | Core distributed processing framework |
| **Python / Matplotlib** | Visualization and reporting |

**Environment Setup:**

* Hadoop 3.x
* Hive 3.x
* Cloudera

**Data Flow Architecture**

**+-------------------+**

**| vgsales.csv |**

**+---------+---------+**

**|**

**v**

**+-------------------+**

**| HDFS |**

**+---------+---------+**

**|**

**v**

**+-------------------+**

**| Hive |**

**+---------+---------+**

**|**

**v**

**+-------------------+**

**| Data Visualization|**

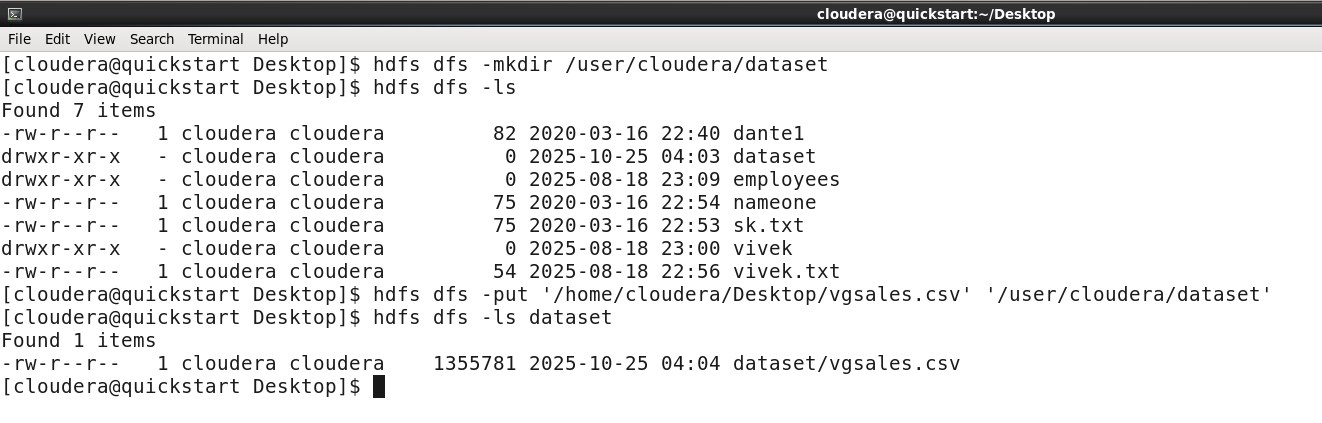
**+-------------------+**

**Performing Analysis on Hive**

Step 1: Upload Data to HDFS

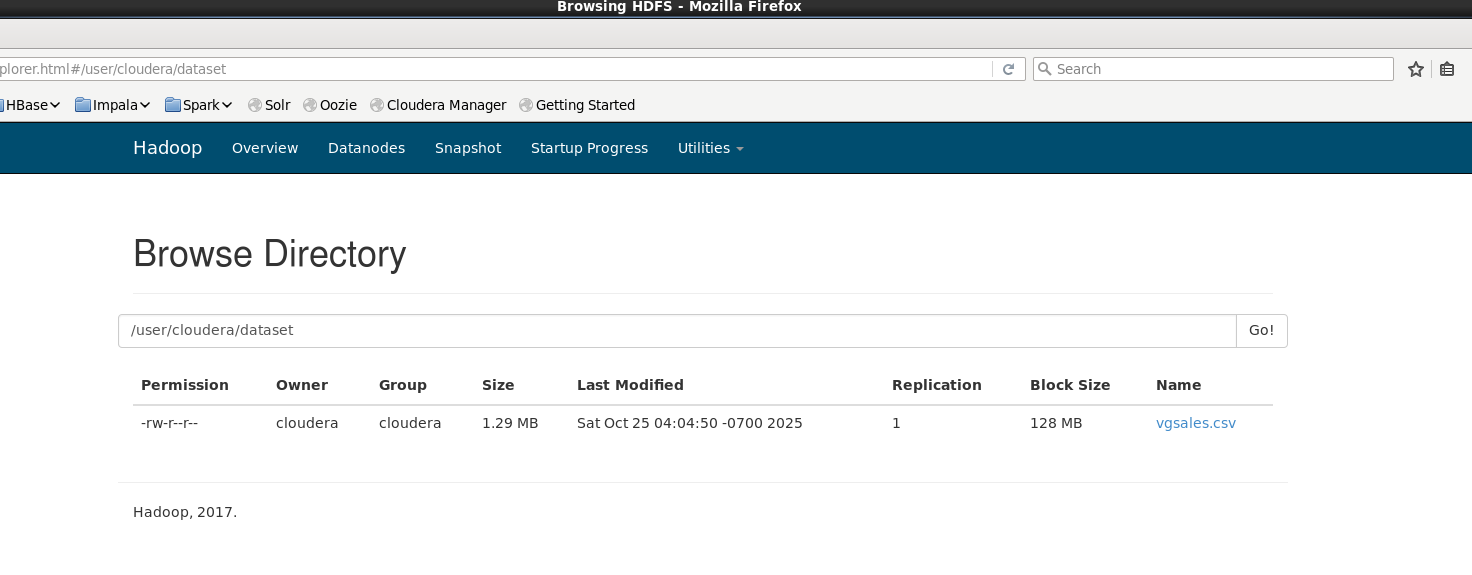
**hdfs dfs -mkdir /user/cloudera/dataset/**

**hdfs dfs -put /home/cloudera/Desktop/vgsales.csv /user/cloudera/dataset**



The dataset appears in our hdfs directory.

We can verify using web interface as well.



Step 2: Create Table in Hive

**CREATE TABLE vgsales (**

**Rank INT,**

**Name STRING,**

**Platform STRING,**

**Year INT,**

**Genre STRING,**

**Publisher STRING,**

**NA\_Sales FLOAT,**

**EU\_Sales FLOAT,**

**JP\_Sales FLOAT,**

**Other\_Sales FLOAT,**

**Global\_Sales FLOAT**

**)**

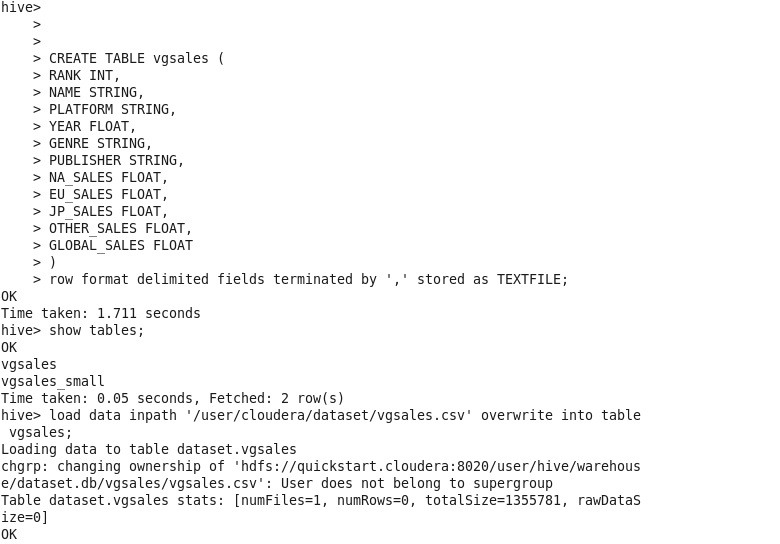
**ROW FORMAT DELIMITED**

**FIELDS TERMINATED BY ','**

**STORED AS TEXTFILE**

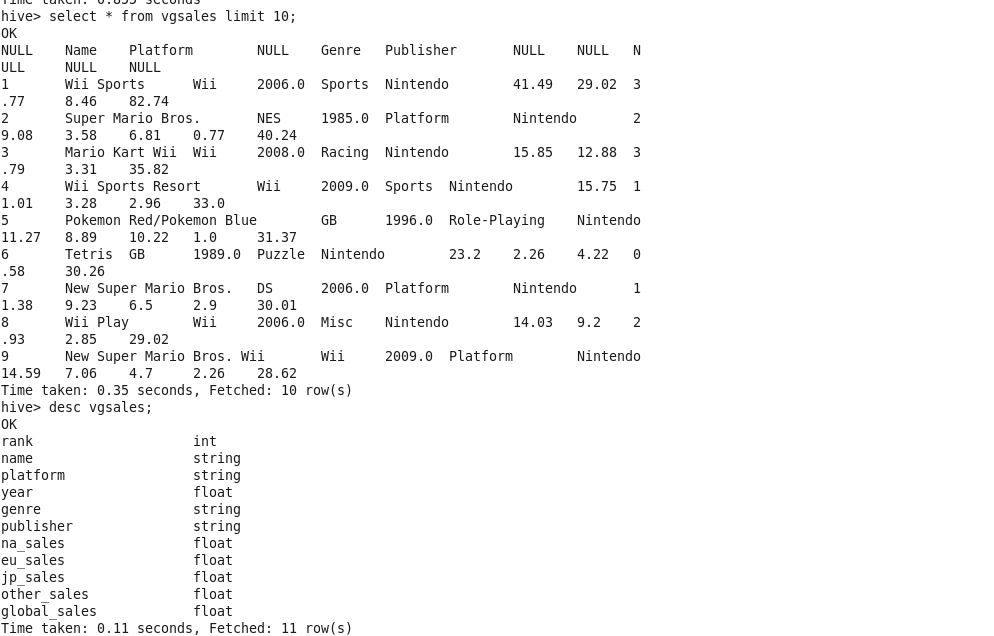
Step 3: Loading data into Table (vgsales) in Hive

**LOAD DATA INPATH ‘/user/cloudera/dataset/vgsales.csv’ OVERWRITE INTO TABLE vgsales**

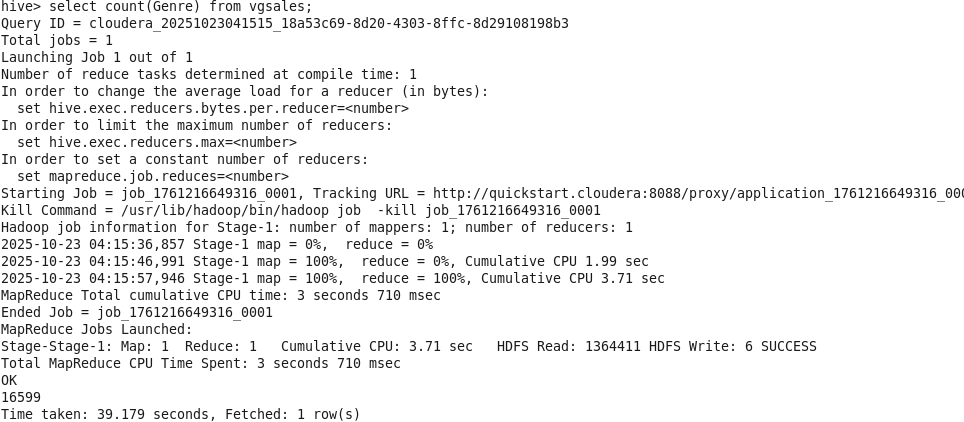
****

Step 4: Performing Analysis / Hive Queries

* Displaying some of the data in hive table



* Getting the total numbers of rows in our hive table.

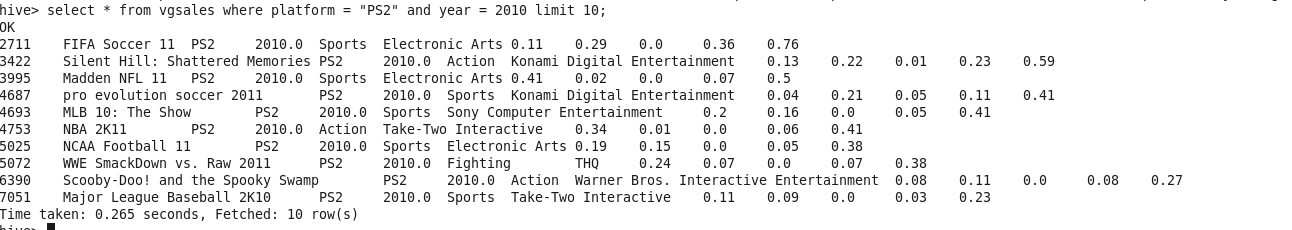


* Getting videos games counts per genre.



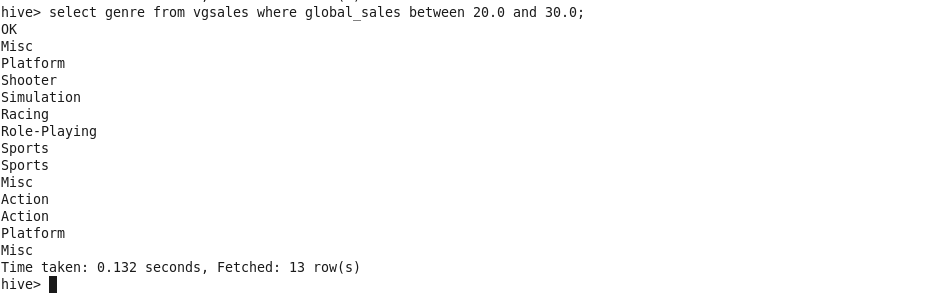
* Displaying titles for specific platform and year.

(helps understand which year has major releases for a specific platform.)



* Displaying popular genres in our video games sales dataset.

(helps understand publisher which genre are popular and smartly invest in those genre.)



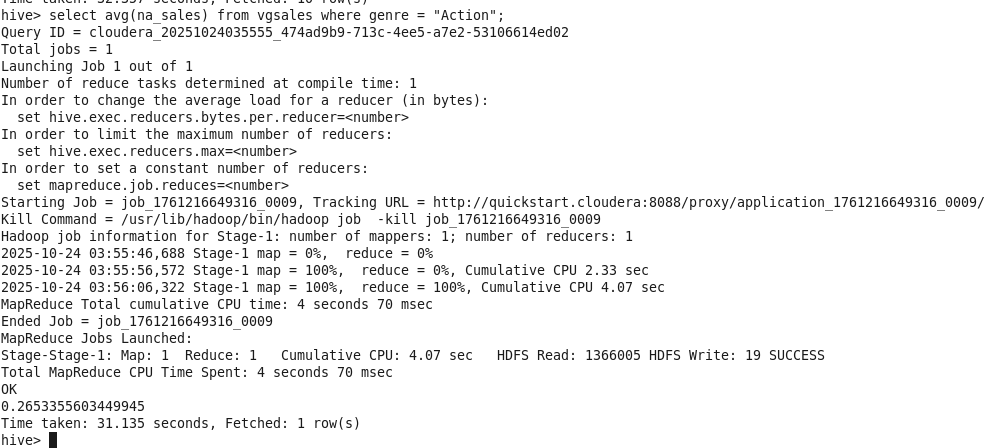
* Displaying genre per publishers.

(helps understand which genre are popular among a publisher.)



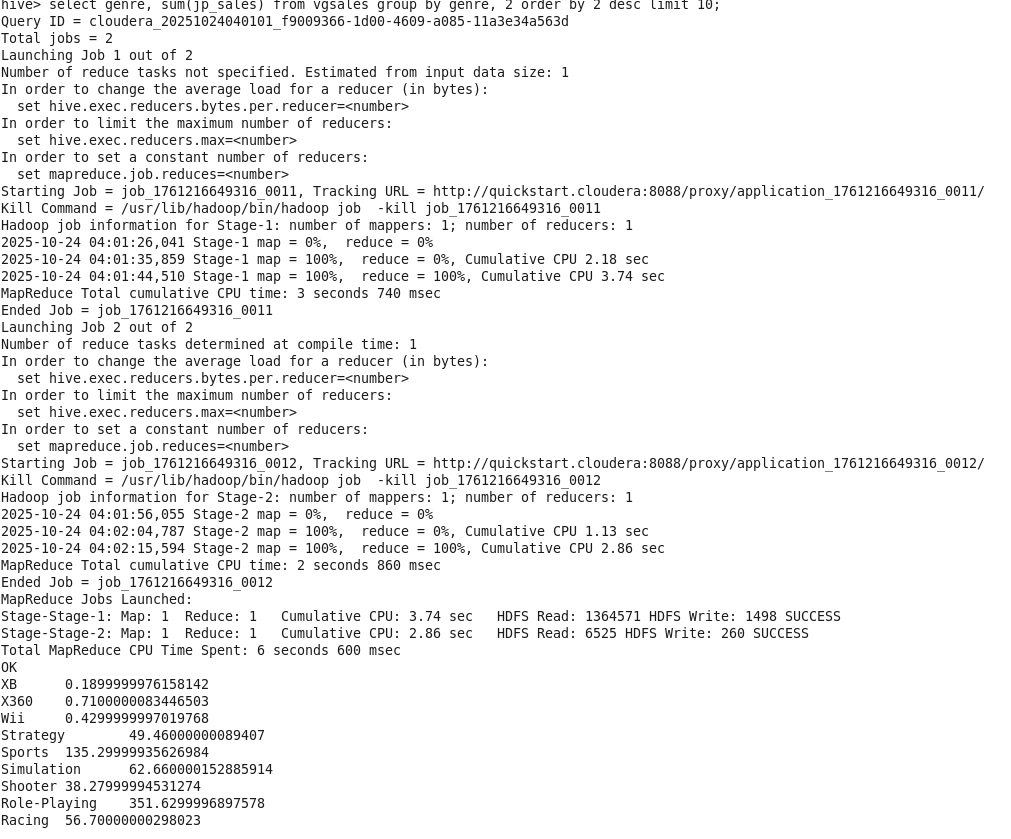
* Displaying average sales in a specific region for a genre.

(helps understand which region to target for maximum sales for a specific genre.)



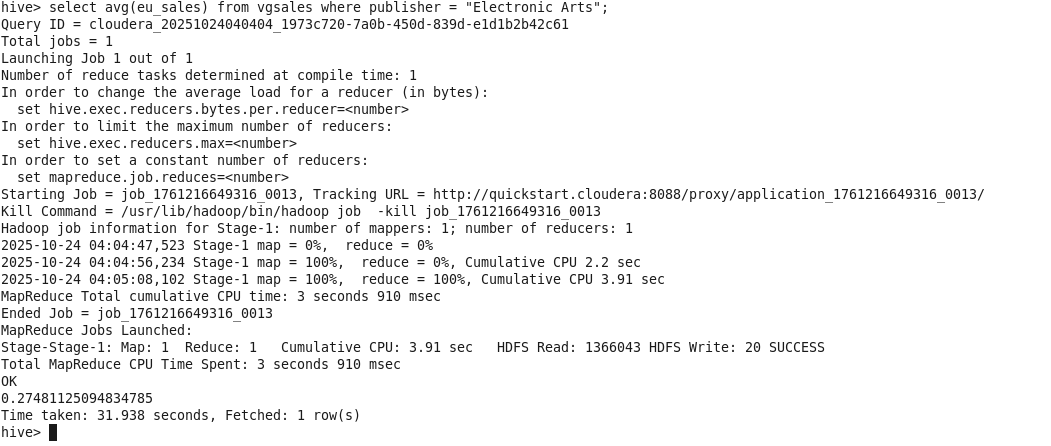
* Displaying total sales of each genre for a region.

(helps understand which genres are more popular in a specific region.)



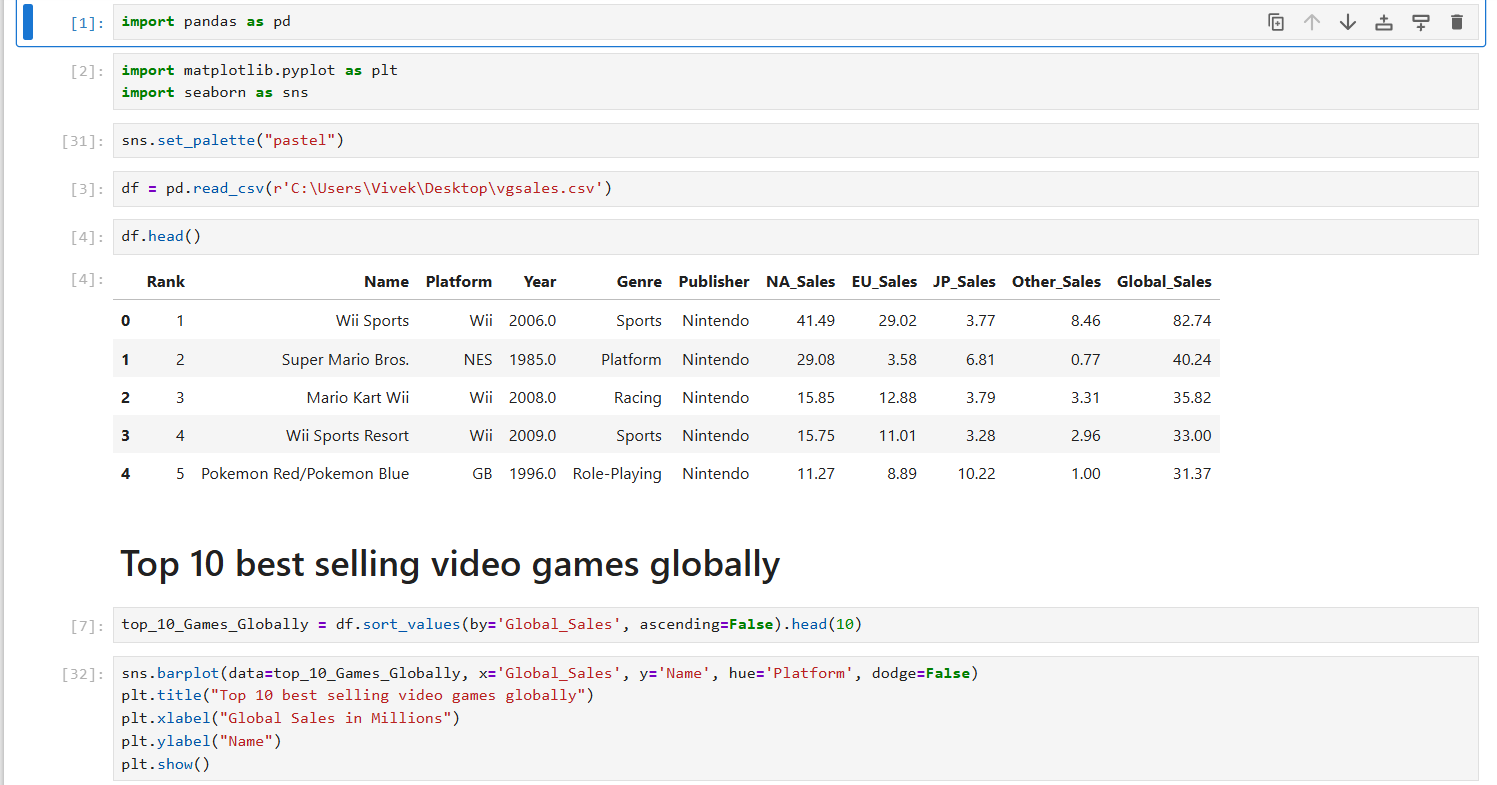
* Displaying average sales of a publisher for a region.

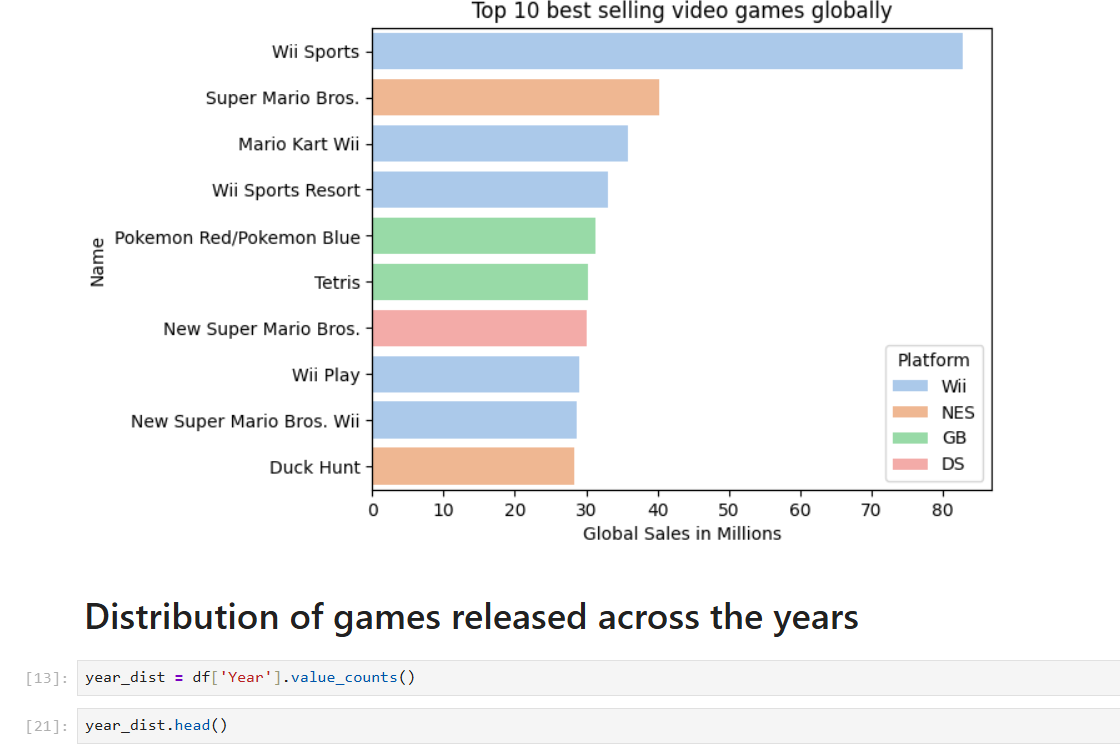
(helps understand a publisher which region to target for maximum sales.)



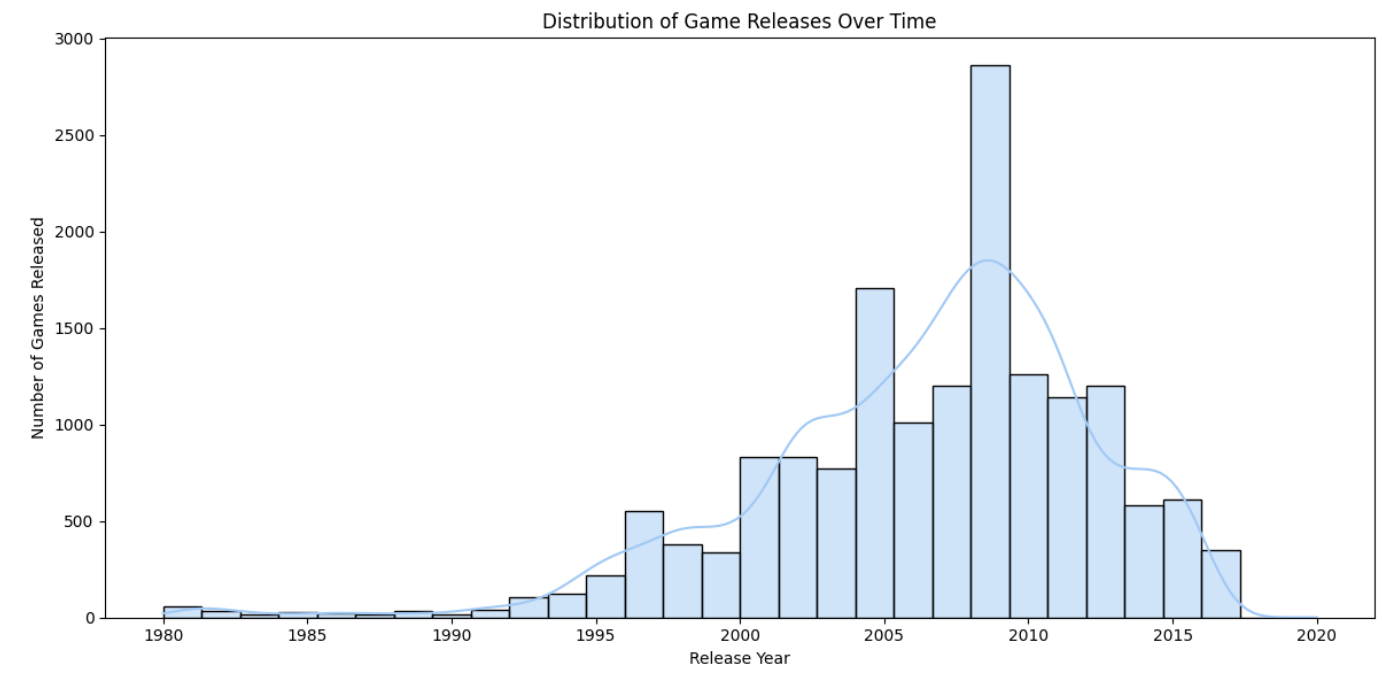
**Data Visualization**

Using Python (Matplotlib / Seaborn):

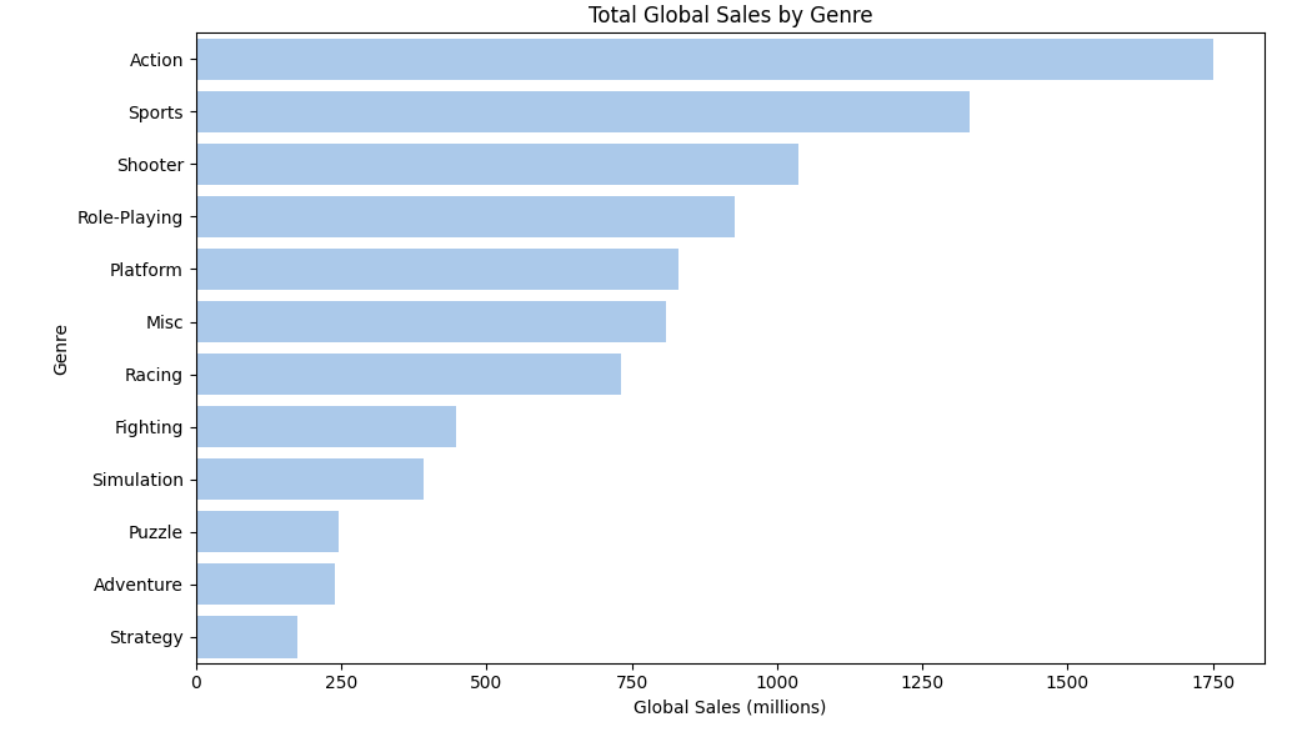




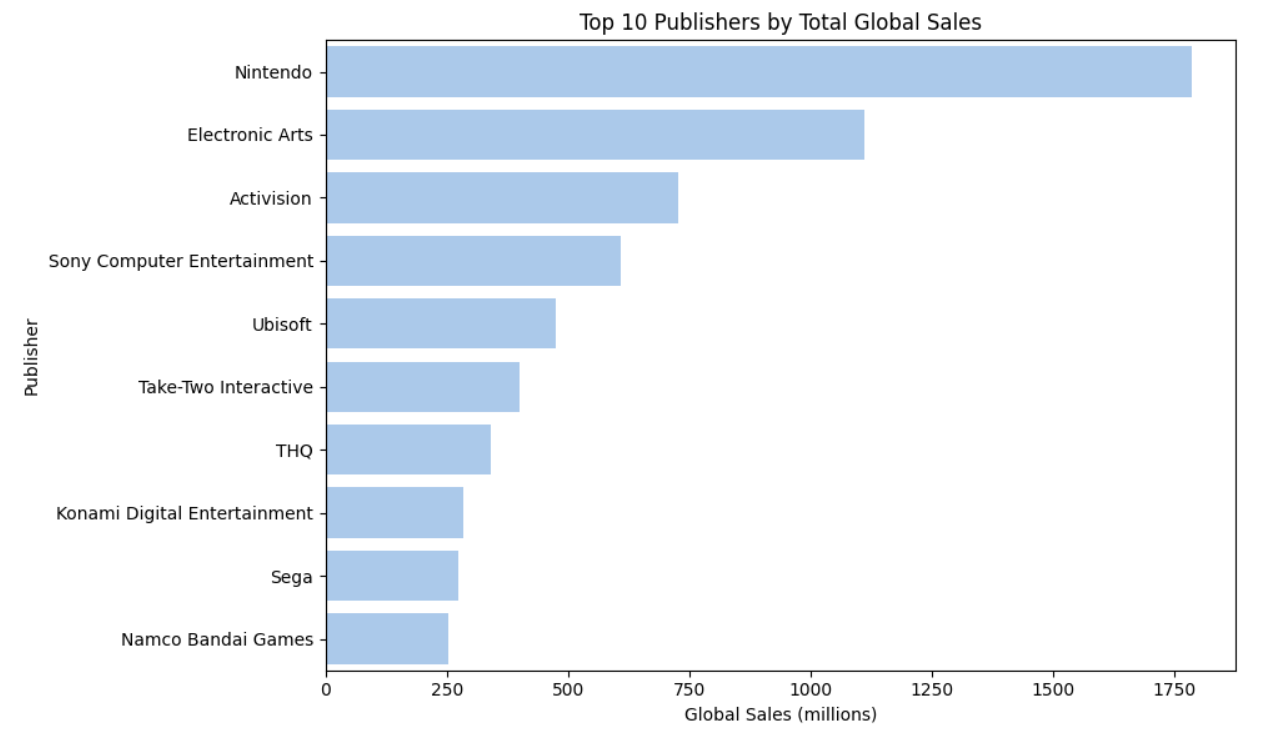


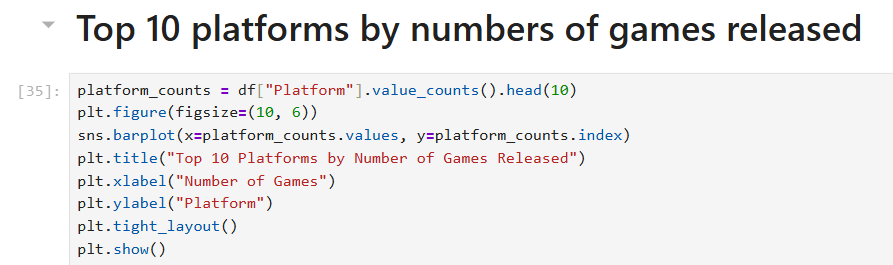


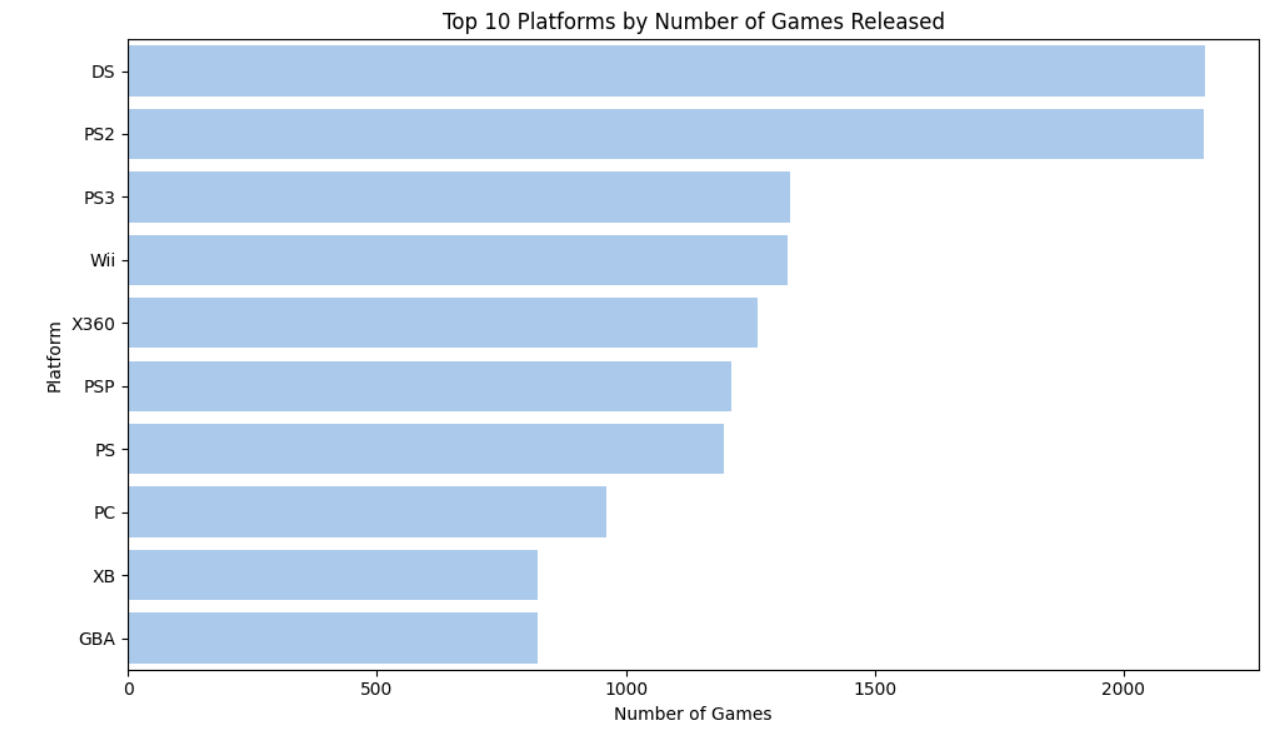


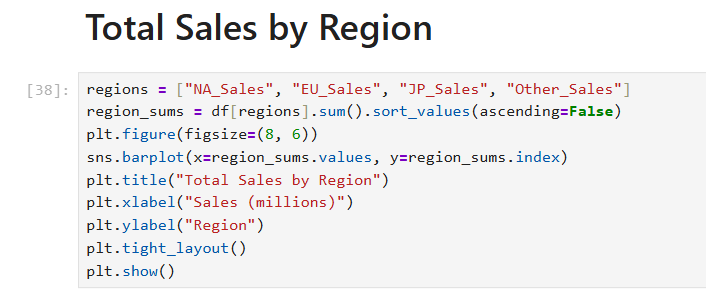


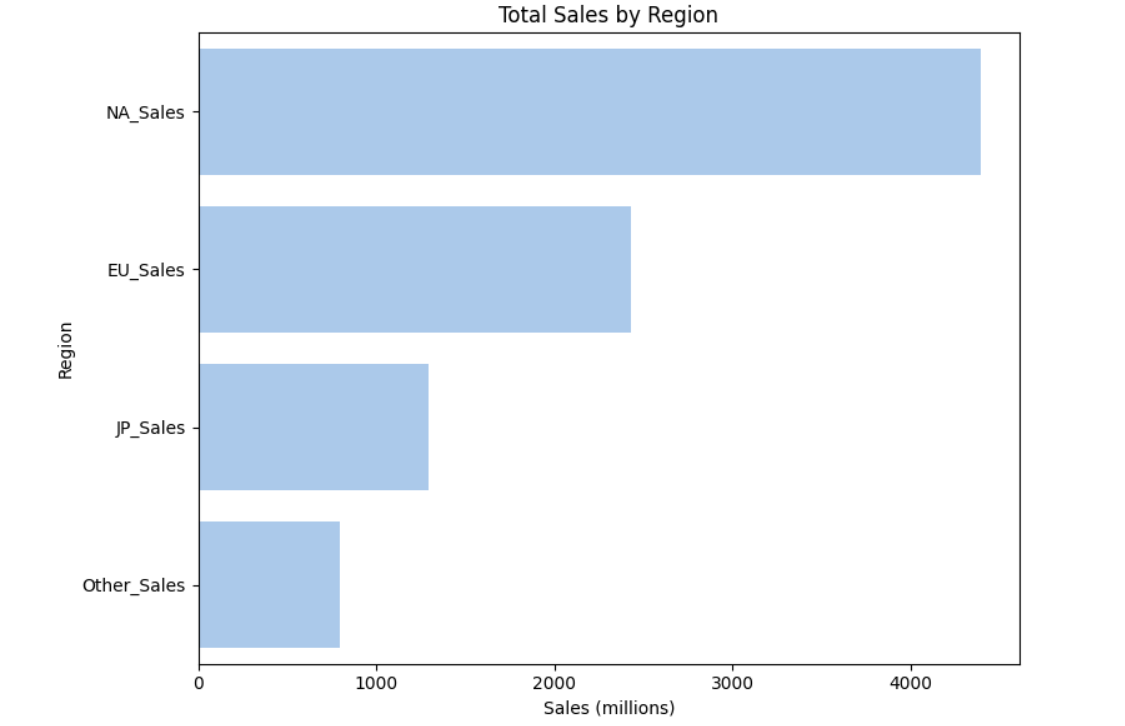


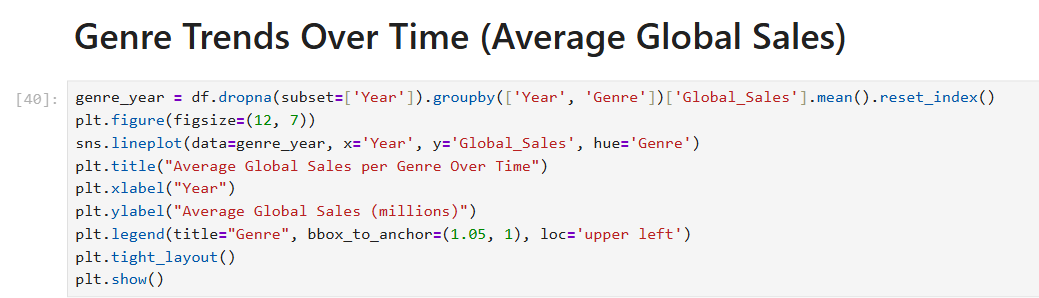


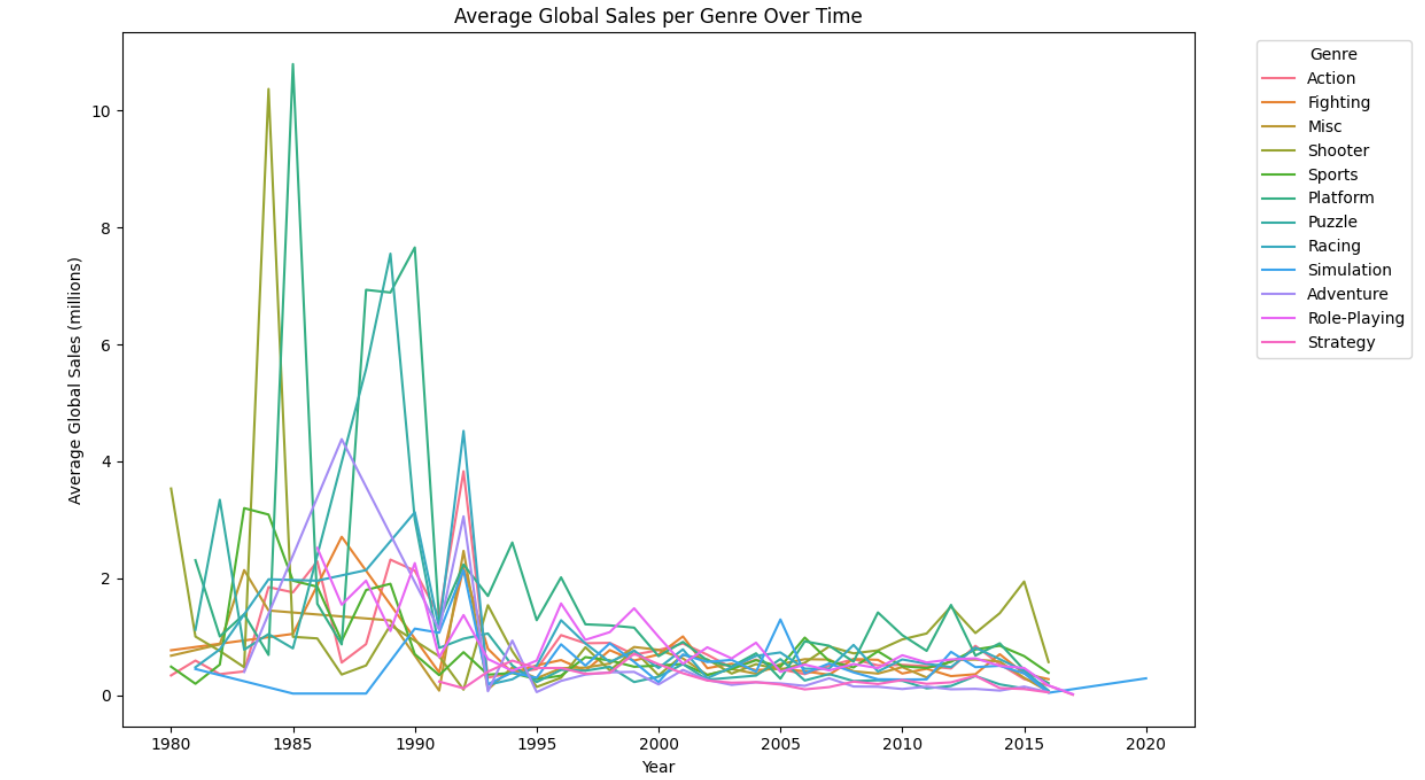


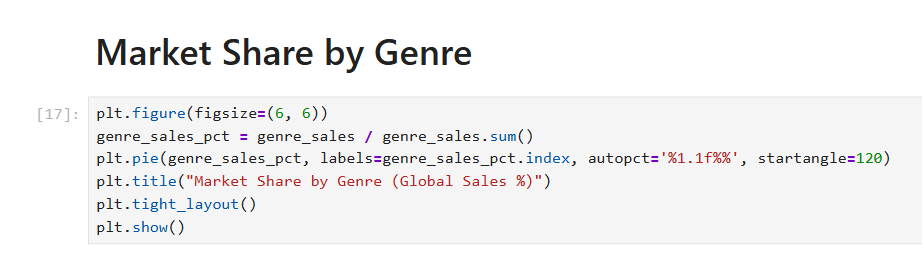














**Conclusion**

This case study showcases how Hadoop ecosystem components — **HDFS**, and **Hive** — can be used to process and analyze structured data at scale.  
Compared to traditional systems like MySQL, **Hive** allows distributed querying, fault tolerance, and better scalability for large datasets.  
Through visualization, we can derive meaningful insights that support business decisions in the gaming industry.