GROUP 6 CAPSTONE PROJECT:

TRENDS IN VIDEO GAME SALES AND A PREDICTIVE ANALYSIS OF THE INDUSTRY

BY: HUGH MCMURRAY, JASON SINGH, RYAN SHELL, & TJ FUGELSETH

THE TEAM

TJ
Graduated from Lake Forest College
BS Physics and Economics

Jason
Graduated from the University of Minnesota (Twin Cities)
MS Mechanical Engineering
BS Physics and Astrophysics

Ryan
Graduated from South Dakota School of Mines
BS Computer Engineering

Hugh
Graduated from the University of Pittsburgh
BS Mathematics

BACKGROUND

- Wanted to examine historical sales of video games
  - Try to make predictive analysis on how much a game would sell given our data
- Main factors included critic score, genre, developer, and more
- Technical Goals: Simulate data stream from the cloud
- Analysis Goals: Find what factors affect video game sales

EXPLORATORY QUESTIONS

1. How does a game's rating impact its sales?
2. Do certain genres sell more than others?
3. How are global sales trending over time?
4. What factors have the most influence on game sales?
5. Do different consoles sell more games than others?
6. How much does a game's content rating impact sales?

TECHNOLOGIES USED

- Kafka
- PowerBI
- Azure Data Lakes
- Azure Databricks
- Azure Data Factories
- Python
- SQL
- Machine Learning

Power BI

SQL

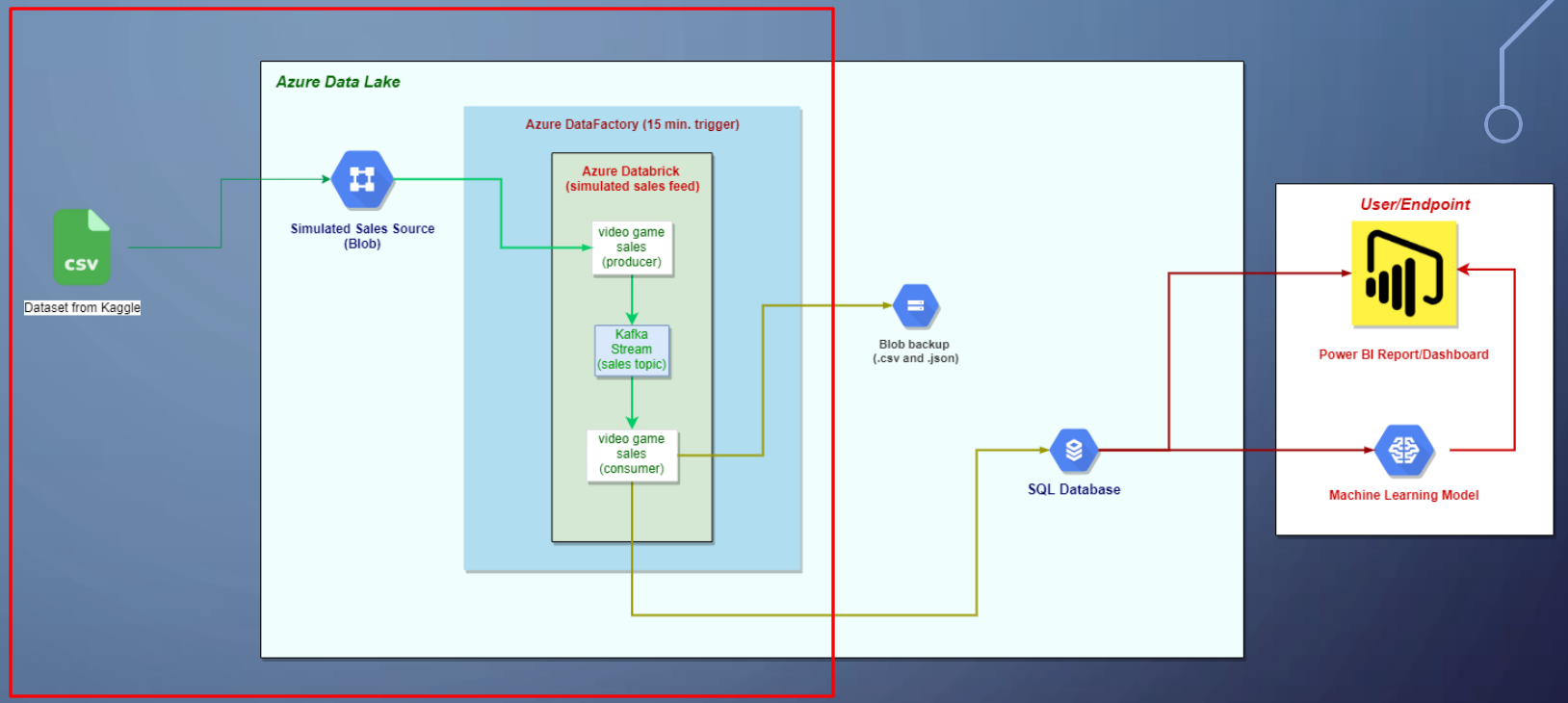
kafka

OUR DATA

- Sources we retrieved data from include:
  - Kaggle
  - VGChartz
  - US Census Bureau

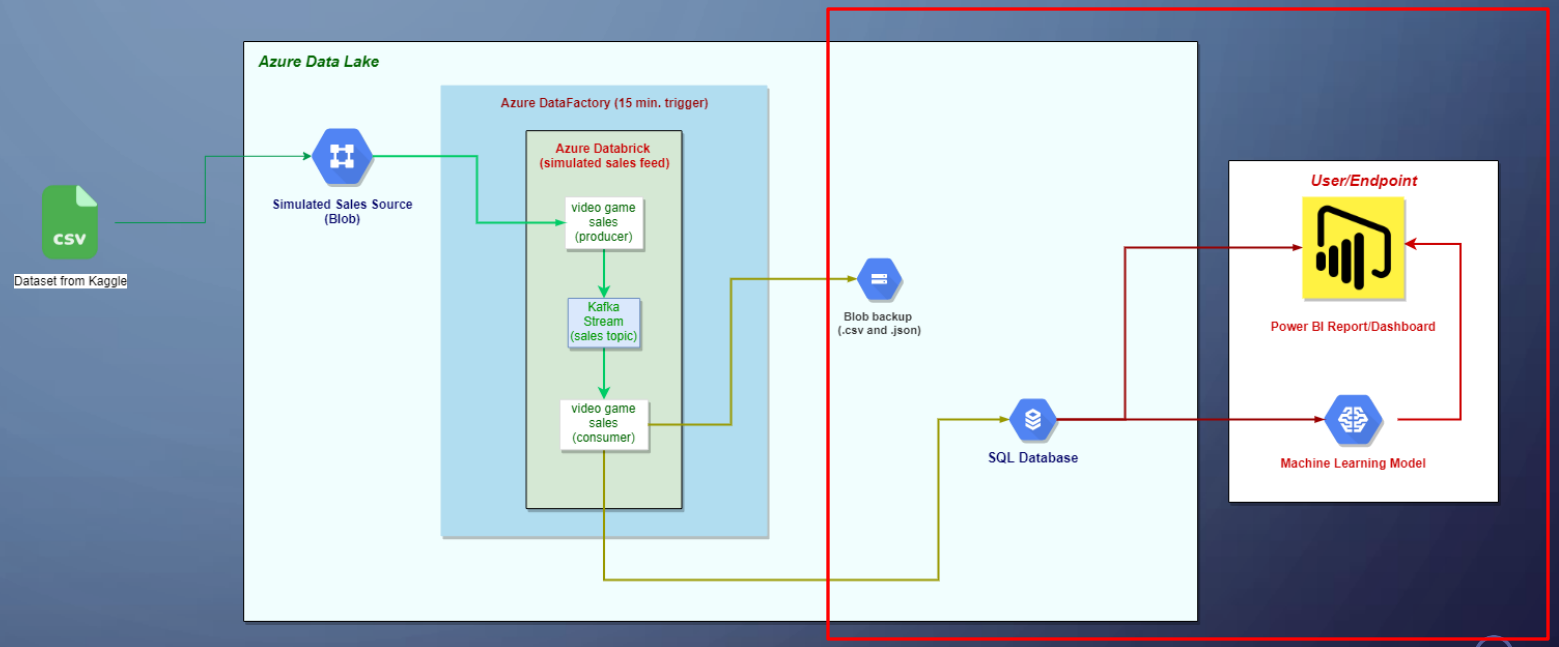
OUR DATA PLATFORM

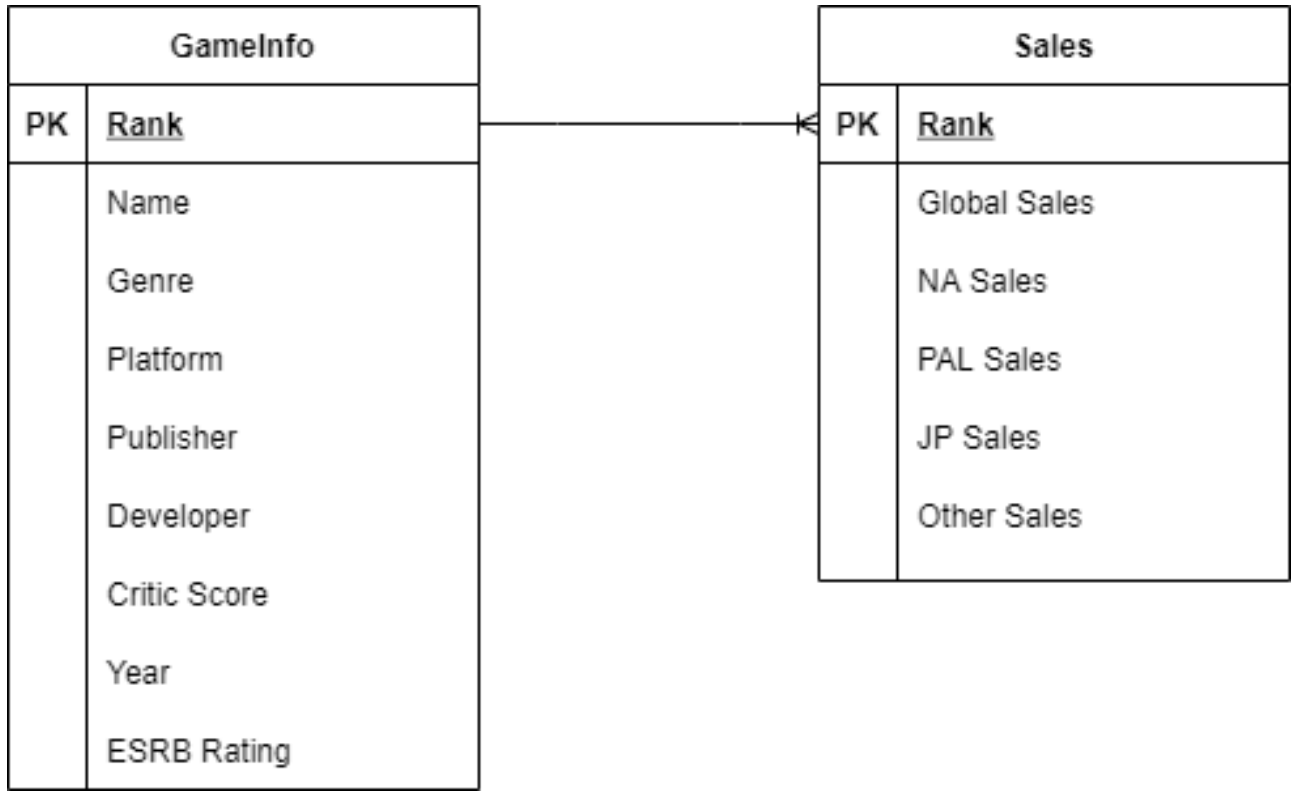
- Our goal is to simulate a DataStream
- Started with manually uploading a csv to the cloud
- In Azure Databricks:
  - Grabbed data from csv in the blob
  - Set up a producer to send messages to the cloud
  - Then set up a consumer to get data
  - Store the consumed data in a blob for backup



DATA PLATFORM (CONTINUED)

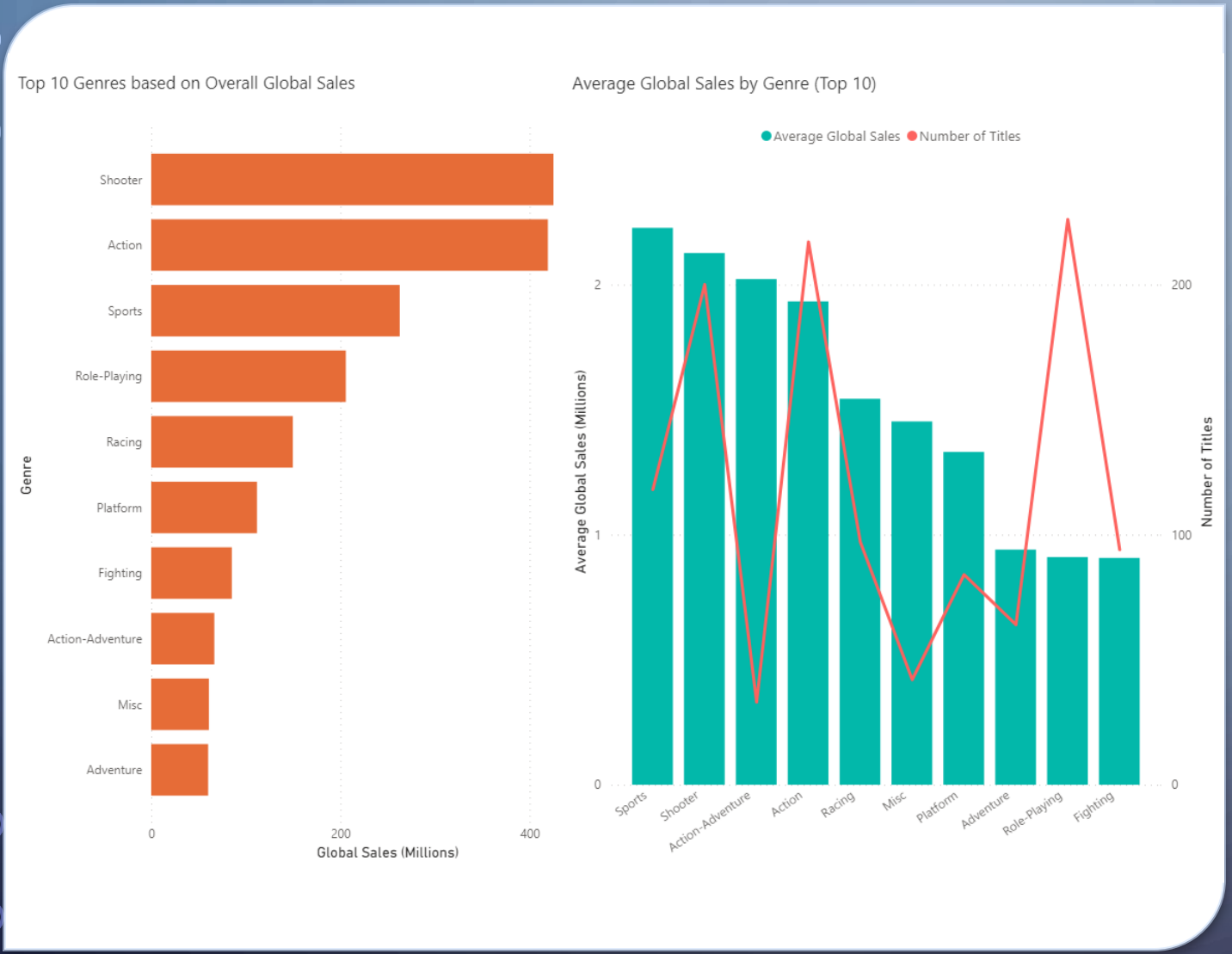
- After data is consumed, send to SQL database
- Once in the SQL Server:
  - The ML model pulls data from the server and sends results to Dashboard
  - The Dashboard pulls data from the server





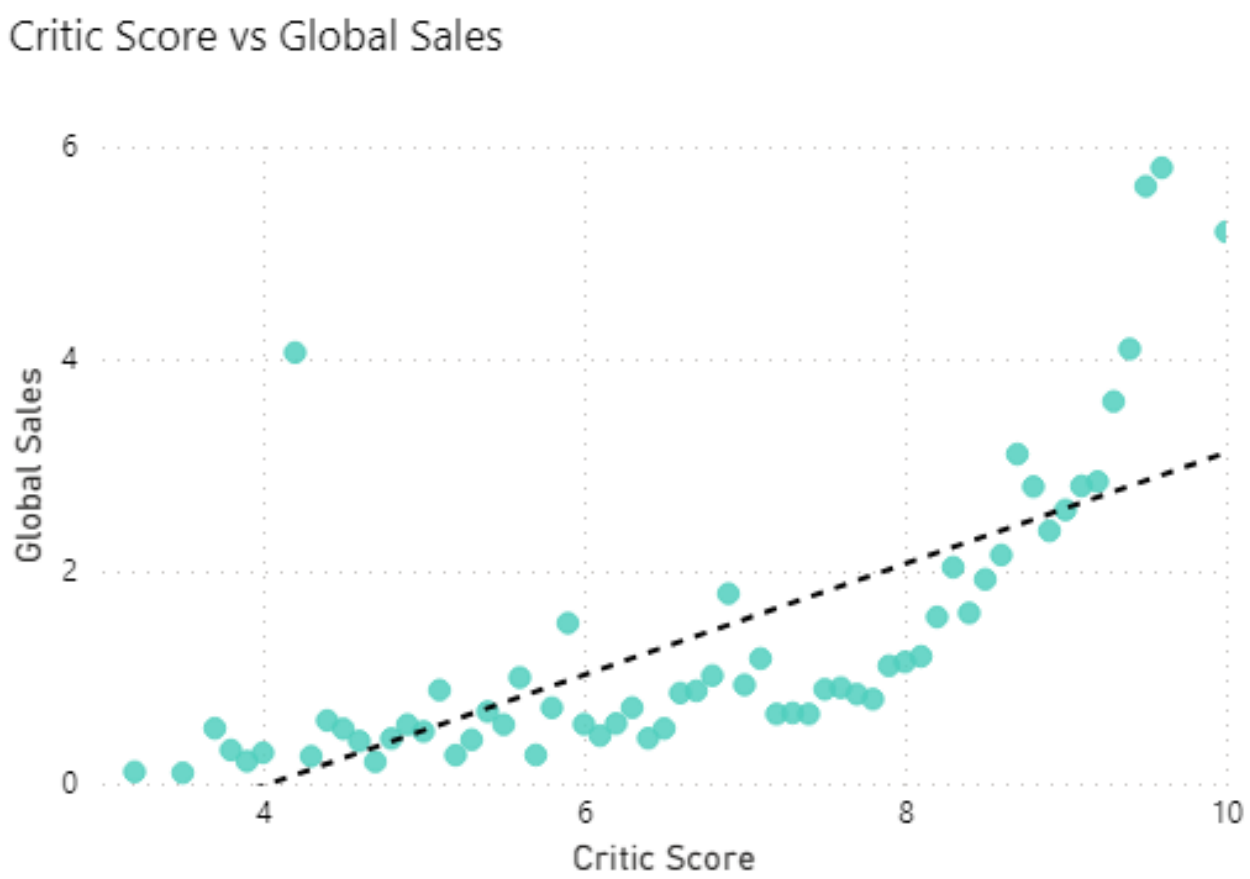
SQL DATABASE

- Temporary table created in SQL database
- DDL Schema used to create tables as shown to the right
- DML Script used to pull data from temporary table into tables created using DDL Schema



RESULTS

- Top Genres based on overall global sales are Shooter, Action and Sports
- Top Genres based on average global sales are Sports, Shooter and Action-Adventure

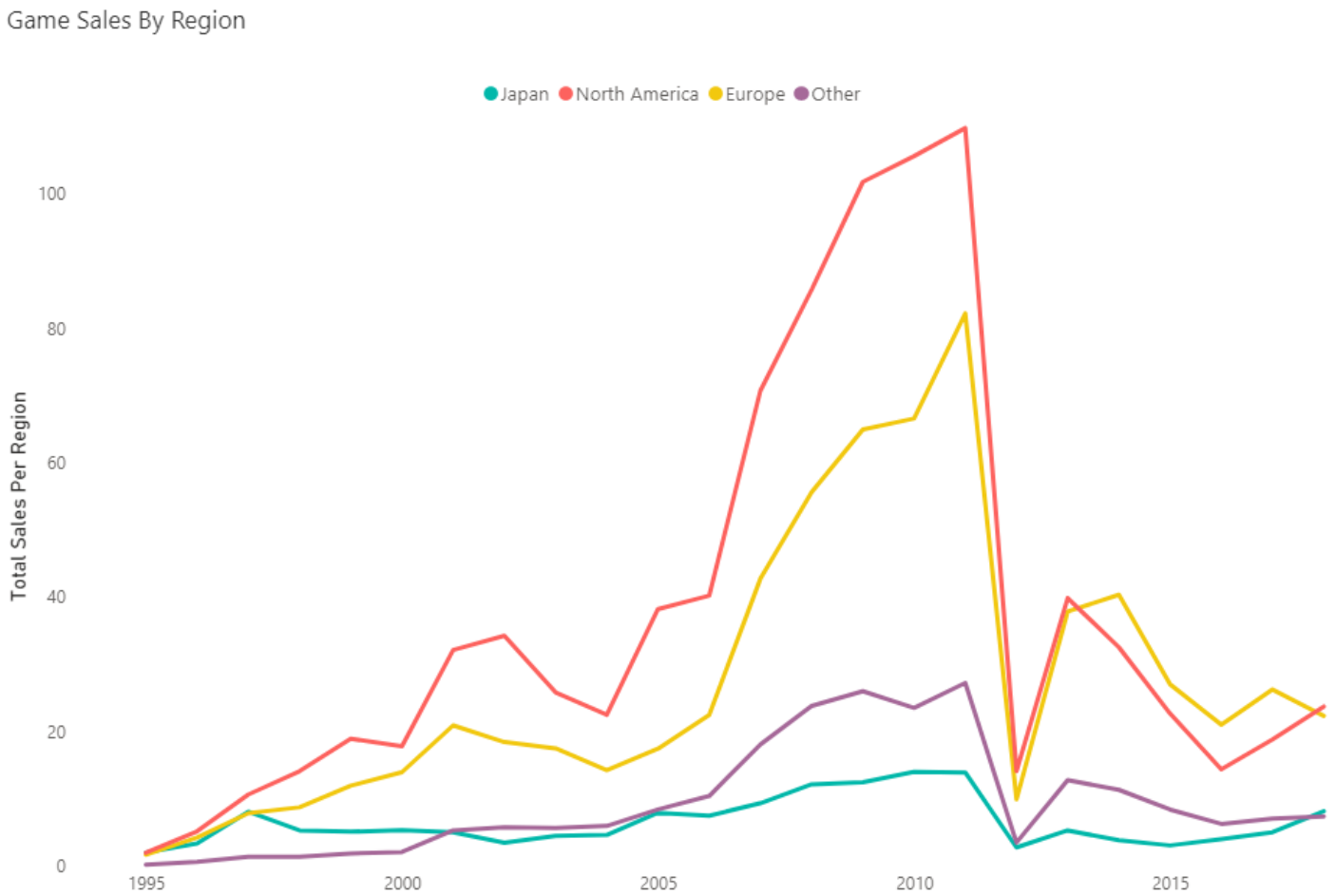


RESULTS

- Scatterplot shows Global Sales vs Critic Score
- As critic score increases so does the sales of a game
- The trendline is upward sloping indicating a positive correlation

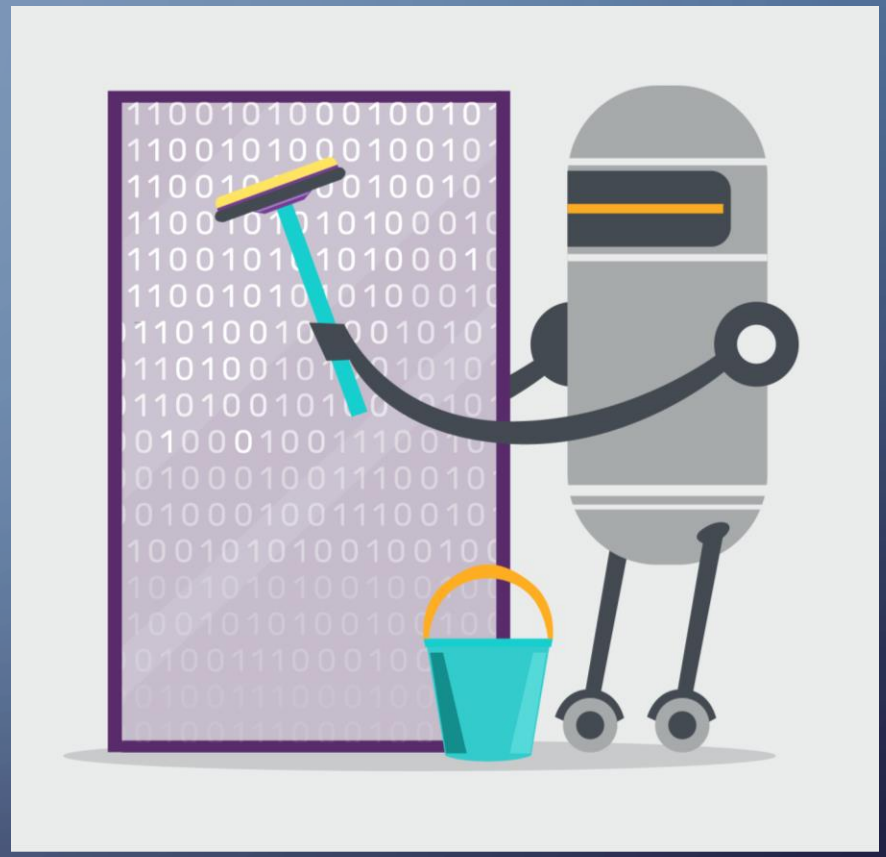
RESULTS

- Line graph shows sales over time for different regions
- North America is almost always the top selling region
- Japan sales tend to always be the lowest
- Sharp drop theorized to be because of changes



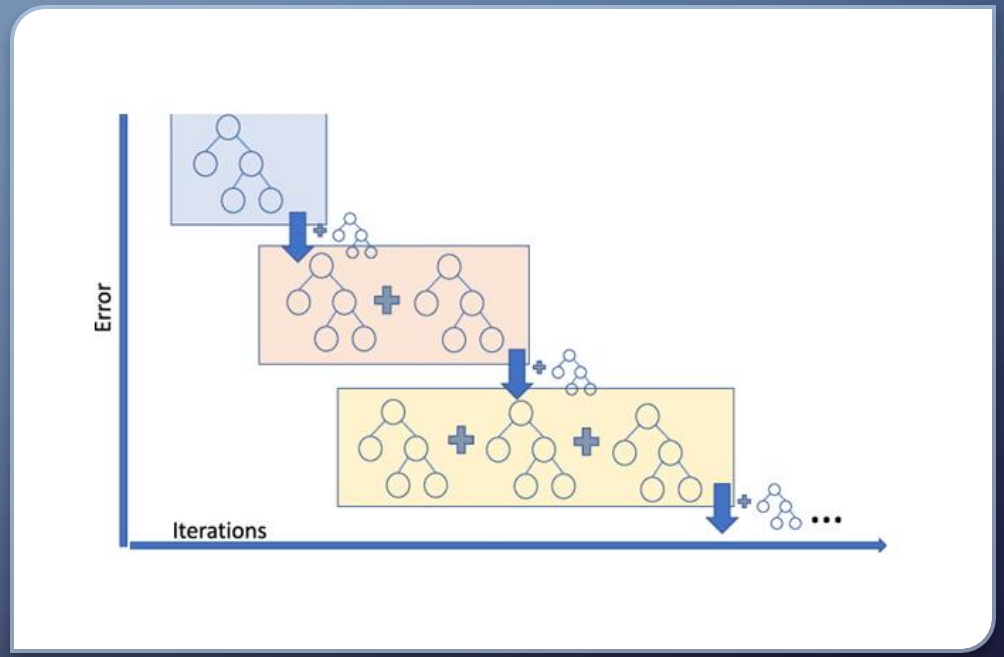
DATA PREPARATION:

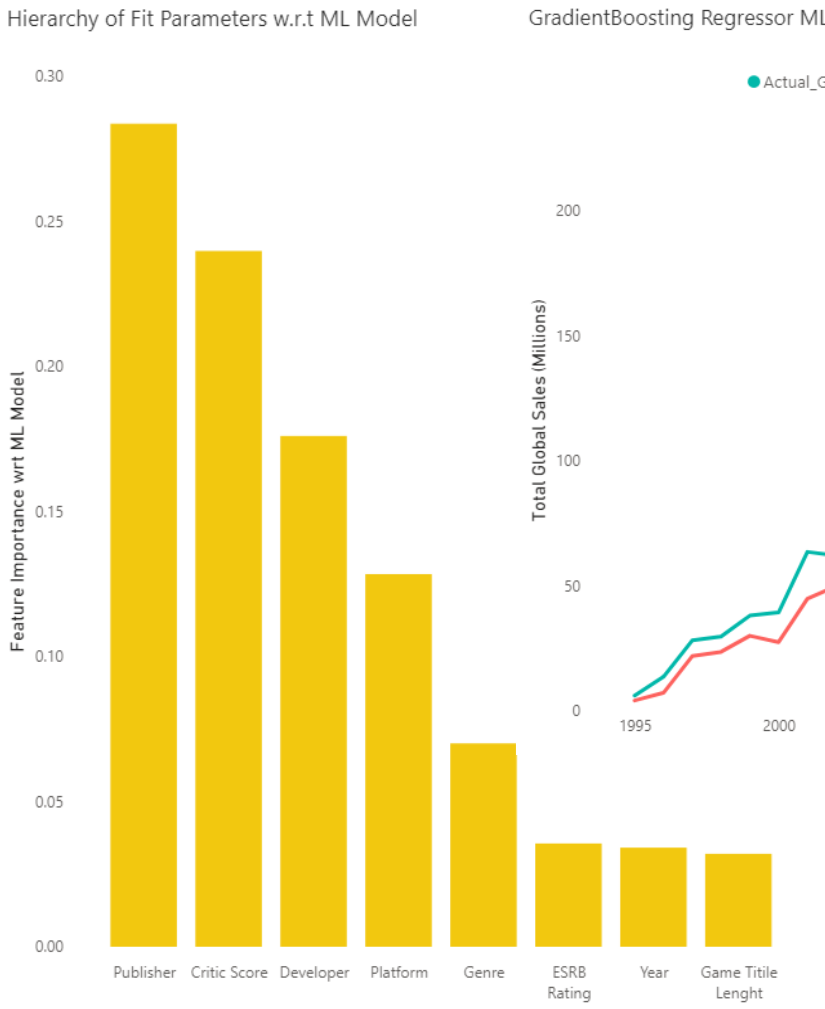
- Removed Nulls
- Grouped Developers and Publishers by prominence in dataset.
- Curious about the influence of a title's name length on its sales.
- Applied log-normal transformation to global sales before feeding it into our model

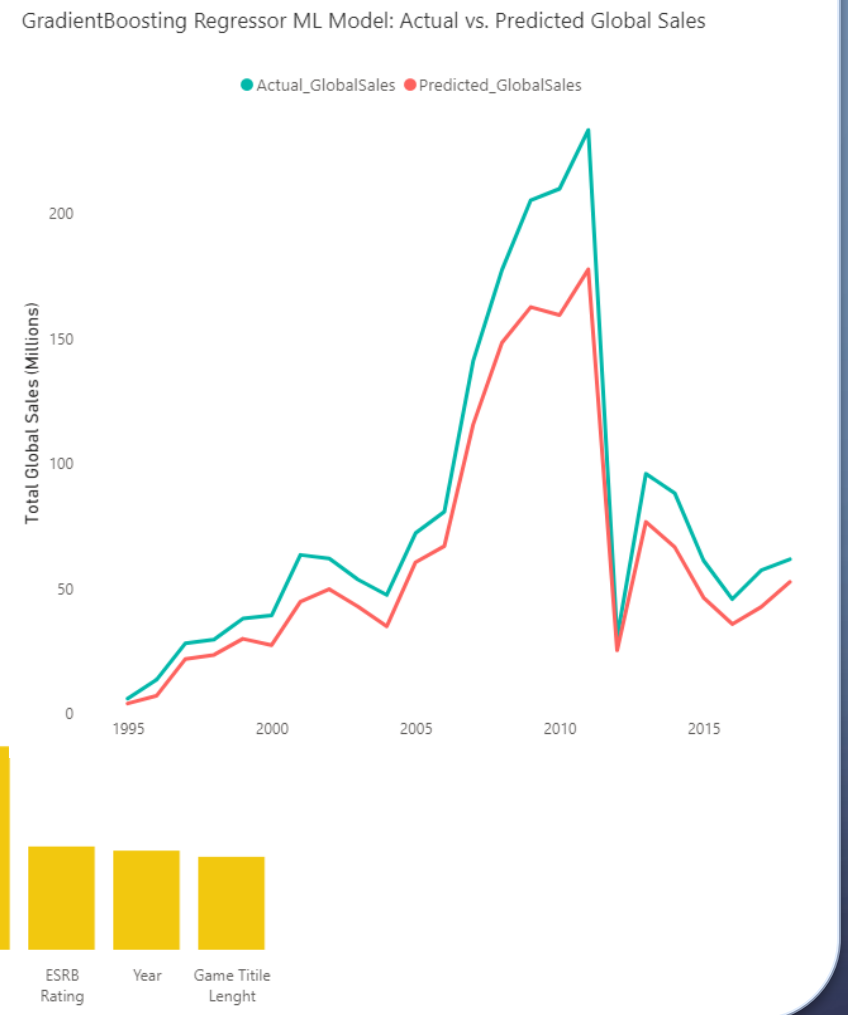


ON THE ML MODEL: GRADIENT BOOSTING REGRESSION

- It is an ensemble technique which uses multiple weak learners to produce a stronger model
- It requires:
  - A loss function to optimize
  - A weak learner to make predictions
  - An additive model to add weak learners to minimize the loss function







MACHINE LEARNING OUTCOMES:

The GradientBoostRegressor() from sklearn's ensemble directory was used. It was optimized using a grid-search to give the following parameters:

- Learning rate: 0.001
- Max depth: 5
- Max features: sqrt
- N estimators: 15,000

After Tuning our model is approx. 60% accurate in its predictive ability

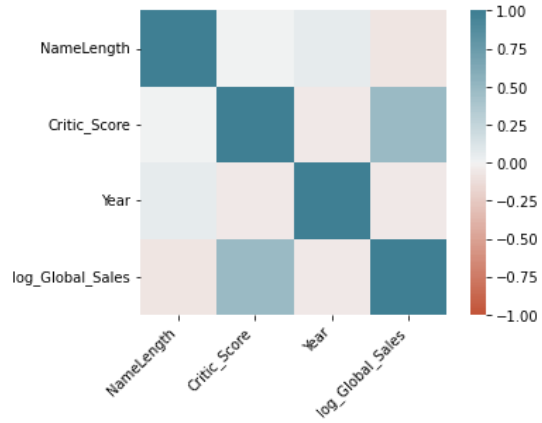
LIVE DASHBOARD DEMO

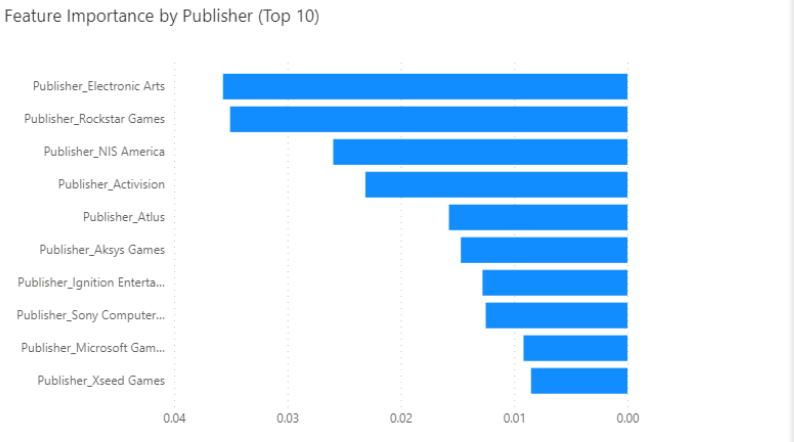
RECOMMENDATIONS FOR IMPROVEMENT

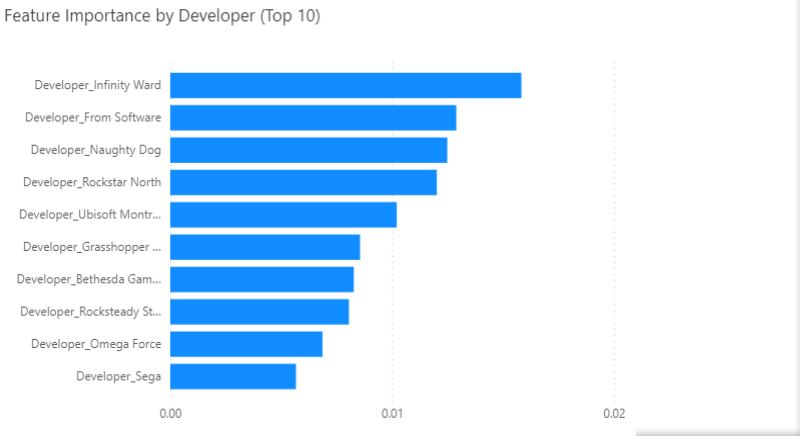
- Finding a cleaner dataset
  - Data had a ton of missing values, was very, very messy
  - Probably the reason we couldn't get above 60% for accuracy
- Ideas:
  - Could try and find digital sales information

SO YOU WANT TO MAKE VIDEO GAME: NOW WHAT?

- Publish with a large/well known publisher
  - Publisher had the highest fit score on the model
- Make an sports or shooter game
  - Both had the highest average sales per game
- Good reviews help too
  - Critic score has a positive correlation with sales, as per the correlation matrix







Data Sources

References:

1.) ***Video Games Sales 2019***, *"Sales and Scores for more than 55,000 games"*. Retrieved from Kaggle

2.) ***Video Game Dataset***, *"474417 Game with Metacritic Score, Ratings, Genres, Publishers, Platforms, ..."*
Retrieved from Kaggle

3.) VGChartz (For web-scraping)

4.) US Census Bureau, Retail Trade: Summary Statistics for the U.S., States, and Selected Geographies: 2017. Survey/Program: Economic Census, TableID: EC1744BASIC, Dataset: ECNBASIC2017. (directlink)

ANY QUESTIONS?

- Thank you!
- Data sources: