Video Game Sales Analysis

•••

Justin Lam

LinkedIn

Portfolio

Goal

The purpose of this project is to take deep dive into analyzing past video game sales and to build a dashboard that will measure the most important KPIs.

The following pages document my thought processes and steps I took in this analysis

Data Sources

Video Game Sales:

The first dataset contains 15,000+ rows of video game sales data scraped from the web alongside partial data including review ratings and maturity ratings.
https://www.kaggle.com/datasets/rush4ratio/video-game-sales-with-ratings

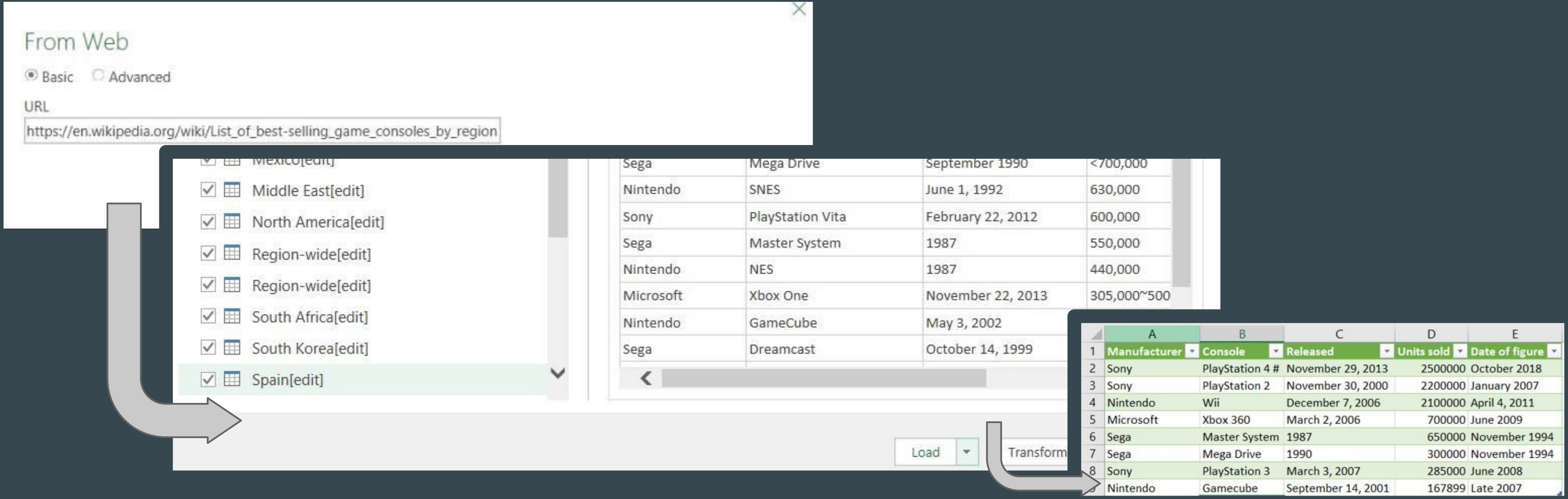
Video Game Console Sales:

The second dataset we use contains various tables of sales data for video game consoles in different regions.

https://en.wikipedia.org/wiki/List_of_best-selling_game_consoles_by_region

Data Collecting - Power Query

Our first data source is already easily obtained in a CSV format, but our 2nd dataset, Video Game Console Sales, comes in the form of already made tables on a website. To scrap this data, we use Power Query and then save the data as a xlsx.



From Web

● Basic ○ Advanced

URL

https://en.wikipedia.org/wiki/List_of_best-selling_game_consoles_by_region

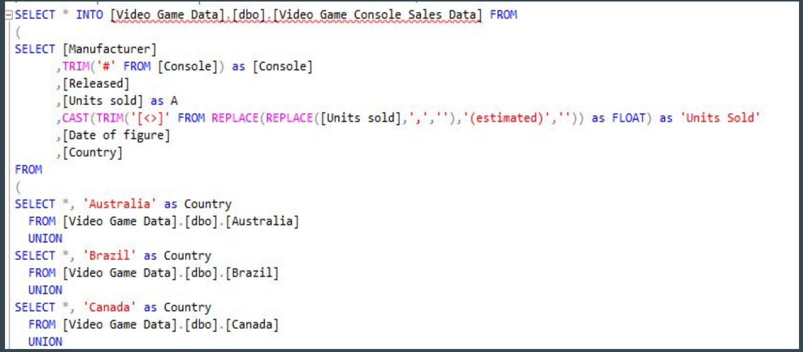
☑ Mexico[edit]
☑ Middle East[edit]
☑ North America[edit]
☑ Region-wide[edit]
☑ Region-wide[edit]
☑ South Africa[edit]
☑ South Korea[edit]
☑ Spain[edit]

Sega	Mega Drive	September 1990	<700,000
Nintendo	SNES	June 1, 1992	630,000
Sony	PlayStation Vita	February 22, 2012	600,000
Sega	Master System	1987	550,000
Nintendo	NES	1987	440,000
Microsoft	Xbox One	November 22, 2013	305,000~500
Nintendo	GameCube	May 3, 2002	
Sega	Dreamcast	October 14, 1999	

Manufacturer	Console	Released	Units sold	Date of figure
Sony	PlayStation 4 #	November 29, 2013	2500000	October 2018
Sony	PlayStation 2	November 30, 2000	2200000	January 2007
Nintendo	Wii	December 7, 2006	2100000	April 4, 2011
Microsoft	Xbox 360	March 2, 2006	700000	June 2009
Sega	Master System	1987	650000	November 1994
Sega	Mega Drive	1990	300000	November 1994
Sony	PlayStation 3	March 3, 2007	285000	June 2008
Nintendo	Gamecube	September 14, 2001	167899	Late 2007

Data Cleaning #1 - SQL Server / T-SQL

After uploading both datasets to SQL Server, one of the first things we notice is our Video Game Console Sales data is divided in individual tables by region, whereas for our analysis, we primarily want to know how well a particular console sold globally.



```sql
SELECT * INTO [Video Game Data].[dbo].[Video Game Console Sales Data] FROM
(
SELECT [Manufacturer]
      ,TRIM('#' FROM [Console]) as [Console]
      ,[Released]
      ,[Units sold] as A
      ,CAST(TRIM('[<>]' FROM REPLACE(REPLACE([Units sold],',',''),'(estimated)','')) as FLOAT) as 'Units Sold'
      ,[Date of figure]
      ,[Country]
FROM
(
SELECT *, 'Australia' as Country
  FROM [Video Game Data].[dbo].[Australia]
  UNION
SELECT *, 'Brazil' as Country
  FROM [Video Game Data].[dbo].[Brazil]
  UNION
SELECT *, 'Canada' as Country
  FROM [Video Game Data].[dbo].[Canada]
  UNION
```

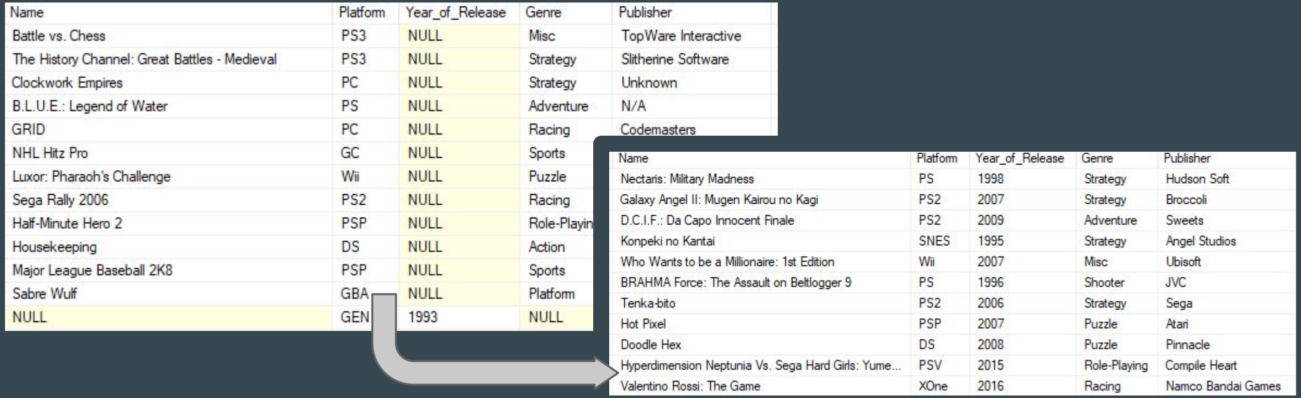
To achieve this we write a SQL query to create a table where we union all regions together, as well as perform some preliminary data cleaning. For example when we look at units sold, we see some values such as “>17,800” which we do not want if we want to be able to aggregate this data.

Link to Full SQL Query

Data Cleaning #2- SQL Server / T-SQL

Afterwards we check for any null data and duplicate rows in both datasets. We write a query to check for this, and then also write follow up queries to delete any unwanted incomplete rows or duplicate rows.

One thing to not is we will keep rows with incomplete reviews/ratings and keep this in mind for future analysis



Name	Platform	Year_of_Release	Genre	Publisher
Battle vs. Chess	PS3	NULL	Misc	TopWare Interactive
The History Channel: Great Battles - Medieval	PS3	NULL	Strategy	Slitherine Software
Clockwork Empires	PC	NULL	Strategy	Unknown
B.L.U.E.: Legend of Water	PS	NULL	Adventure	N/A
GRID	PC	NULL	Racing	Codemasters
NHL Hitz Pro	GC	NULL	Sports	
Luxor: Pharaoh's Challenge	Wii	NULL	Puzzle	
Sega Rally 2006	PS2	NULL	Racing	
Half-Minute Hero 2	PSP	NULL	Role-Playin	
Housekeeping	DS	NULL	Action	
Major League Baseball 2K8	PSP	NULL	Sports	
Sabre Wulf	GBA	NULL	Platform	
NULL	GEN	1993	NULL	

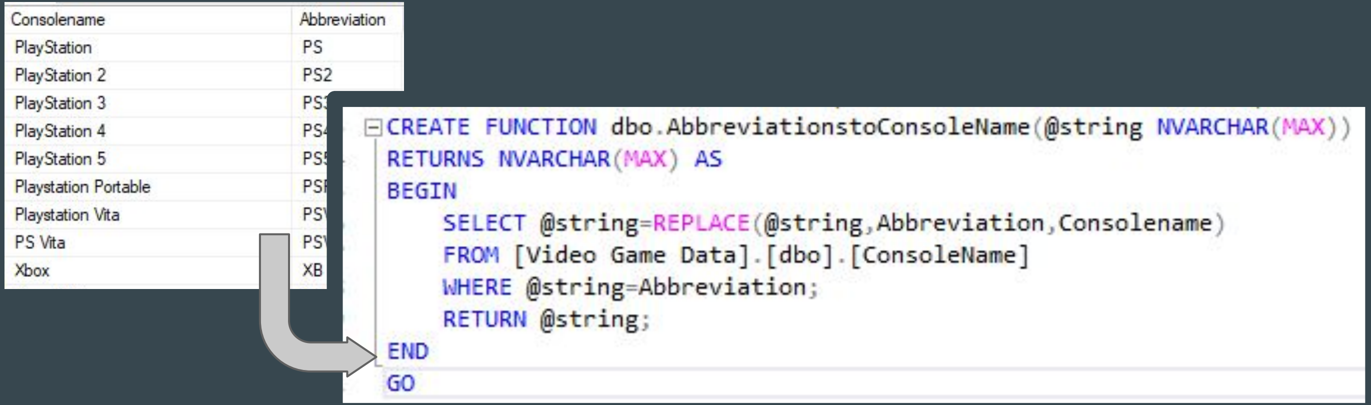
Name	Platform	Year_of_Release	Genre	Publisher
Nectaris: Military Madness	PS	1998	Strategy	Hudson Soft
Galaxy Angel II: Mugen Kairou no Kagi	PS2	2007	Strategy	Broccoli
D.C.I.F.: Da Capo Innocent Finale	PS2	2009	Adventure	Sweets
Konpeki no Kantai	SNES	1995	Strategy	Angel Studios
Who Wants to be a Millionaire: 1st Edition	Wii	2007	Misc	Ubisoft
BRAHMA Force: The Assault on Beltlogger 9	PS	1996	Shooter	JVC
Tenka-bito	PS2	2006	Strategy	Sega
Hot Pixel	PSP	2007	Puzzle	Atari
Doodle Hex	DS	2008	Puzzle	Pinnacle
Hyperdimension Neptunia Vs. Sega Hard Girls: Yume...	PSV	2015	Role-Playing	Compile Heart
Valentino Rossi: The Game	XOne	2016	Racing	Namco Bandai Games

Link to Full SQL Query

Data Cleaning #3- SQL Server / T-SQL

Another problem we run into is to be able join our two datasets we need a shared key but our Video Game Sales dataset has a column with abbreviated console names, whereas our Video Game Console Sales has a column with the full console name.

The solution is to create a new table with the matched abbreviations and full name, and to create a function to replace all abbreviations with the full console name



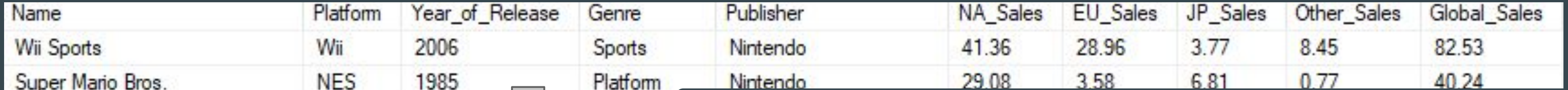
Consolename	Abbreviation
PlayStation	PS
PlayStation 2	PS2
PlayStation 3	PS3
PlayStation 4	PS4
PlayStation 5	PS5
Playstation Portable	PSP
Playstation Vita	PSV
PS Vita	PSV
Xbox	XB

```sql
CREATE FUNCTION dbo.AbbreviationstoConsoleName(@string NVARCHAR(MAX))
RETURNS NVARCHAR(MAX) AS
BEGIN
    SELECT @string=REPLACE(@string,Abbreviation,Consolename)
    FROM [Video Game Data].[dbo].[ConsoleName]
    WHERE @string=Abbreviation;
    RETURN @string;
END
GO
```

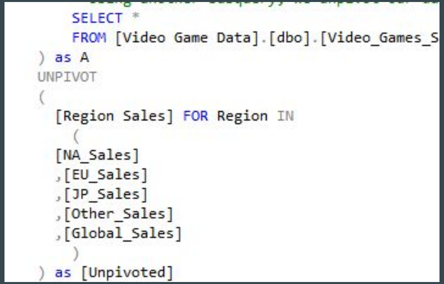
Link to Full SQL Query

Data Cleaning #4- SQL Server / T-SQL

Taking a deeper look we see that our Video Game Sales dataset has regional sales data split into NA, EU, JP, and Other, with a Global (total) column. To be able to analyze this data properly we need to transform the data with each row having its own regional sales data. To do this we write a query to unpivot the data.

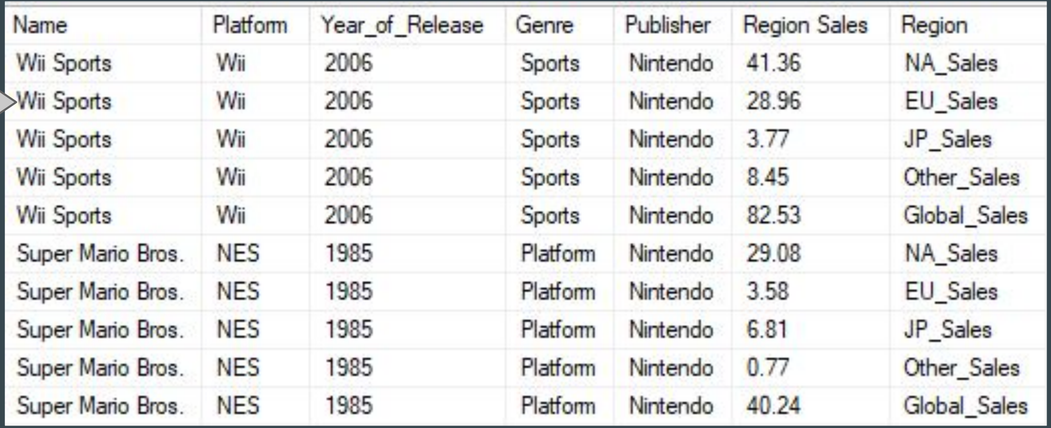


Name	Platform	Year_of_Release	Genre	Publisher	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Global_Sales
Wii Sports	Wii	2006	Sports	Nintendo	41.36	28.96	3.77	8.45	82.53
Super Mario Bros.	NES	1985	Platform	Nintendo	29.08	3.58	6.81	0.77	40.24



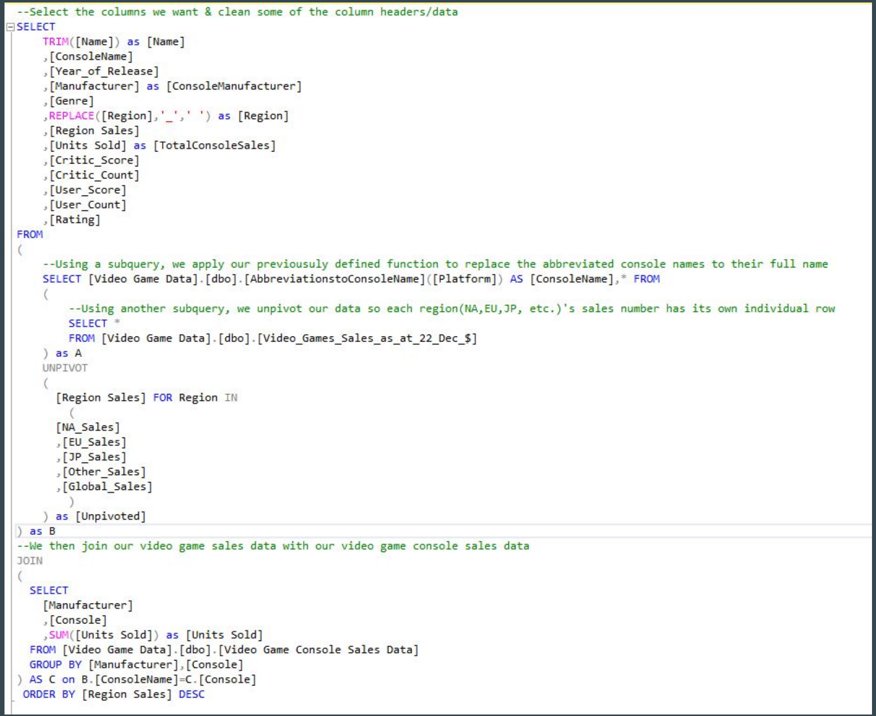
```
      SELECT *
      FROM [Video Game Data].[dbo].[Video_Games_S
  ) as A
  UNPIVOT
  (
    [Region Sales] FOR Region IN
      (
    [NA_Sales]
    ,[EU_Sales]
    ,[JP_Sales]
    ,[Other_Sales]
    ,[Global_Sales]
      )
  ) as [Unpivoted]
```

Link to Full SQL Query



Name	Platform	Year_of_Release	Genre	Publisher	Region Sales	Region
Wii Sports	Wii	2006	Sports	Nintendo	41.36	NA_Sales
Wii Sports	Wii	2006	Sports	Nintendo	28.96	EU_Sales
Wii Sports	Wii	2006	Sports	Nintendo	3.77	JP_Sales
Wii Sports	Wii	2006	Sports	Nintendo	8.45	Other_Sales
Wii Sports	Wii	2006	Sports	Nintendo	82.53	Global_Sales
Super Mario Bros.	NES	1985	Platform	Nintendo	29.08	NA_Sales
Super Mario Bros.	NES	1985	Platform	Nintendo	3.58	EU_Sales
Super Mario Bros.	NES	1985	Platform	Nintendo	6.81	JP_Sales
Super Mario Bros.	NES	1985	Platform	Nintendo	0.77	Other_Sales
Super Mario Bros.	NES	1985	Platform	Nintendo	40.24	Global_Sales

Data Cleaning #3 & 4- SQL Server / T-SQL



```sql
--Select the columns we want & clean some of the column headers/data
SELECT
    TRIM([Name]) as [Name]
    ,[ConsoleName]
    ,[Year_of_Release]
    ,[Manufacturer] as [ConsoleManufacturer]
    ,[Genre]
    ,REPLACE([Region],'_',' ') as [Region]
    ,[Region Sales]
    ,[Units Sold] as [TotalConsoleSales]
    ,[Critic_Score]
    ,[Critic_Count]
    ,[User_Score]
    ,[User_Count]
    ,[Rating]
FROM
(
    --Using a subquery, we apply our previousuly defined function to replace the abbreviated console names to their full name
    SELECT [Video Game Data].[dbo].[AbbreviationstoConsoleName]([Platform]) AS [ConsoleName],* FROM
    (
        --Using another subquery, we unpivot our data so each region(NA,EU,JP, etc.)'s sales number has its own individual row
        SELECT *
        FROM [Video Game Data].[dbo].[Video_Games_Sales_as_at_22_Dec_$]
    ) as A
    UNPIVOT
    (
      [Region Sales] FOR Region IN
        (
      [NA_Sales]
      ,[EU_Sales]
      ,[JP_Sales]
      ,[Other_Sales]
      ,[Global_Sales]
        )
    ) as [Unpivoted]
) as B
--We then join our video game sales data with our video game console sales data
JOIN
(
  SELECT
    [Manufacturer]
    ,[Console]
    ,SUM([Units Sold]) as [Units Sold]
  FROM [Video Game Data].[dbo].[Video Game Console Sales Data]
  GROUP BY [Manufacturer],[Console]
) AS C on B.[ConsoleName]=C.[Console]
 ORDER BY [Region Sales] DESC
```

We combine all previous queries to create a query that ultimately outputs the data we want to analyze and visualize further.

We then export this data into Tableau for further analysis

Link to Full SQL Query

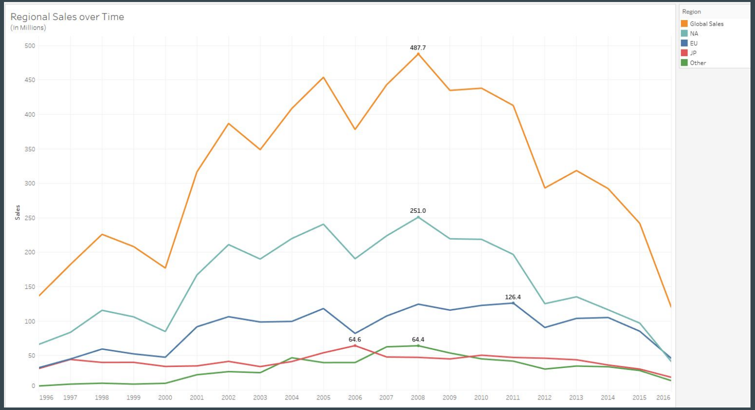
Data Analysis #1- Tableau

Our first analysis is to look at how various video game sales trended over time by region (Global, NA, EU, JP, Other).

Insights:

- The highest global and NA sales were in 2008

- Economic events such as the 2008 Housing Bubble might of impacted sales as we see major declines following 2008



Link to Final Dashboard

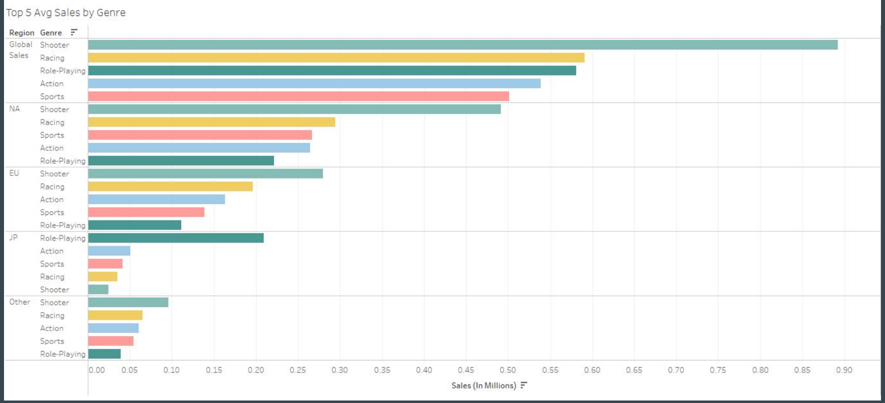
Data Analysis #2- Tableau

Here we wanted to analyze how well do different genres perform in different markets

Insights:

- On average Shooters gross the highest amount of sales in each region except Japan, where Role-Playing games gross the highest on average

- All regions share the same top 5 genres in sales whereas we do not see other genres even listed such as Puzzles or Strategy



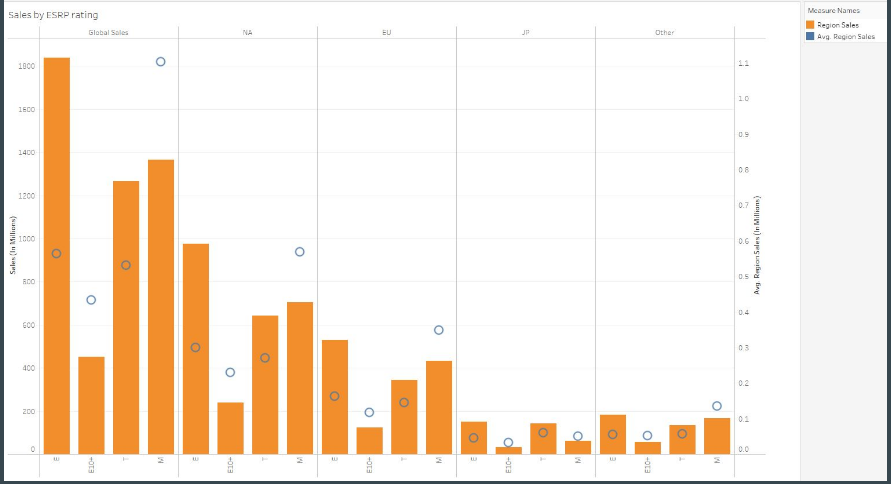
Link to Final Dashboard

Data Analysis #3- Tableau

Next we wanted to analyze how a game is rated (in terms of maturity) effects the sales in video games in different markets. Since we do not have full data on some of the video game's ratings, we are only able to do an analysis on titles we have full information on

Insights:

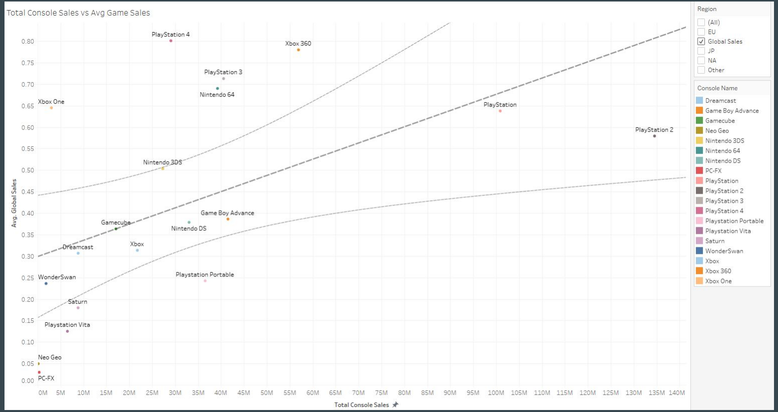
- Generally while "E" rated games have the highest sum sales, it is actually "M" rated games that tend to have the highest sales on average



Link to Final Dashboard

Data Analysis #4- Tableau

Diving deeper, and utilizing data from both datasets, we plot how well a console performs in terms of sales against the average sales performance for video games for that console



Total Console Sales vs Avg Game Sales

Region

☐ (All)
☐ EU
☑ Global Sales
☐ JP
☐ NA
☐ Other

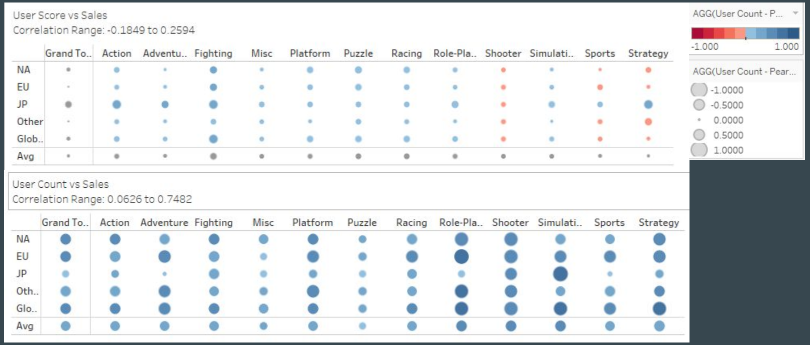
Insights:

- As a particular console sells better, the average game for that console also sells better. Specifically we have a p-value of .018 (which indicates this is statistically significant) and a correlation coefficient of 0.5347 (which indicated a moderate relationship)

Link to Final Dashboard

Data Analysis #5- Tableau

Leveraging the CORR function in tableau, we next create (4) correlation matrices between the critic/user review scores & volume of reviews compared to net sales a video game achieves



User Score vs Sales
Correlation Range: -0.1849 to 0.2594

User Count vs Sales
Correlation Range: 0.0626 to 0.7482

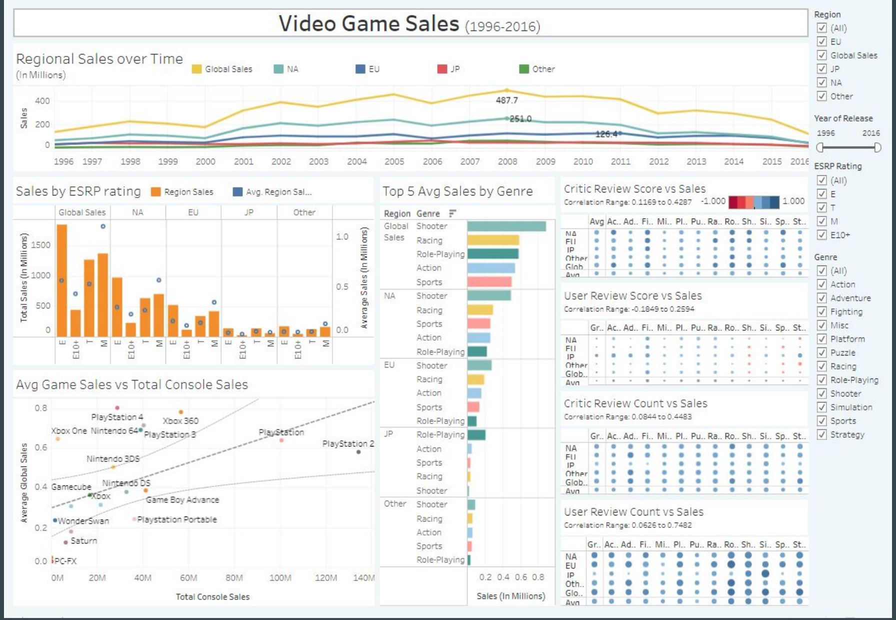
Insights:

- At a high level, the most interesting thing we see is that there are weak (and even some negative) relationships between the the score a user gives and sales, but surprisingly the stronger relationship is tied to the net amount of reviews from users. This backs up the saying “No such thing as bad publicity”.

Link to Final Dashboard

Data Visualization - Tableau

Lastly, by combining our previous sheets and analysis, and with some formatting, we create our final visualization dashboard



Video Game Sales (1996-2016)

Regional Sales over Time (In Millions)

Sales by ESRP rating

Top 5 Avg Sales by Genre

Critic Review Score vs Sales
Correlation Range: 0.1169 to 0.4287

User Review Score vs Sales
Correlation Range: -0.1849 to 0.2594

Critic Review Count vs Sales
Correlation Range: 0.0844 to 0.4483

User Review Count vs Sales
Correlation Range: 0.0626 to 0.7482

Avg Game Sales vs Total Console Sales

Region
☑ (All)
☑ EU
☑ Global Sales
☑ JP
☑ NA
☑ Other

Year of Release
1996 2016

ESRP Rating
☑ (All)
☑ E
☑ T
☑ M
☑ E10+

Genre
☑ (All)
☑ Action
☑ Adventure
☑ Fighting
☑ Misc
☑ Platform
☑ Puzzle
☑ Racing
☑ Role-Playing
☑ Shooter
☑ Simulation
☑ Sports
☑ Strategy

Link to Final Dashboard

End Result

In the end, we joined data from two different sources which we cleaned within SQL Server and created a visualization in Tableau

Our final dashboard lets us filter different fields such as Region, Ratings and Genres to empower ourselves with the information needed to make they correct decisions moving forward