Creating a Profitable Video Game Title: An Analysis Using SQL

by Emerson Fleming

July 14th, 2024

Introduction

Problem:

1. You are a rising video game developer hoping to create the most profitable title of the year. How will you solve this problem?
    a. You are given 2 datasets to build your findings upon. This is to simulate the reality that clean and optimal data will not always be available in a real setting.

        i. **game_sales:** A dataset containing regional video games sales for each title included as well as the overall rating given by Metacritic.
            1. Columns:
                img
                title
                console
                genre
                publisher
                developer
                critic_score
                total_sales
                na_sales
                jp_sales
                pal_sales
                other_sales
                release_date
                last_update

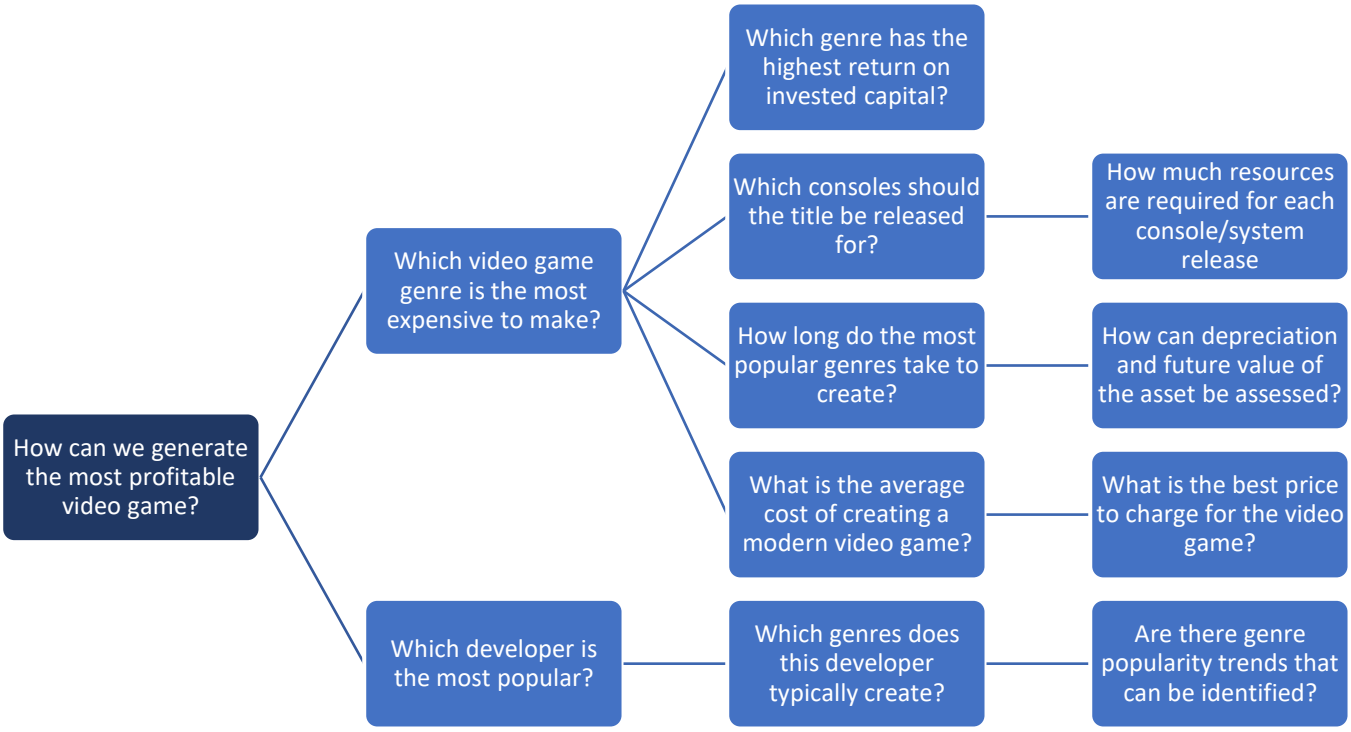
        ii. **game_info:** A dataset containing information from 475,000 titles.
            1. Columns:
                slug
                name
                metacritic
                released
                tba
                updated
                suggestions_count
                platforms
                developers
                genres
                publishers
                esrb_rating

Hypothesis:
The most profitable video game to create relative to cost is an indie survival PC title.

Null Hypothesis:
The most profitable video game to create relative to cost is not an indie survival PC title.

Initial brainstorming questions:

1. Which video game genre is the most expensive to create?
    a. Which games have the highest profit to cost ratio?
2. Which genres have the highest sales?
3. Which platforms should the game be made for?
4. Does the name of the title impact sales?
    a. If so, in what way?
5. Which developer is the most popular?
    a. What genres does this developer typically create?
    b. What are their games generally rated?
    c. Which developer has the best reviews?
        i. What kinds of video games does this developer typically make?
6. Is it feasible to consider release date?
    a. When should the game be released to generate the most sales?
7. Can genre popularity trends be identified over time?
8. Which video games require the most time to create?



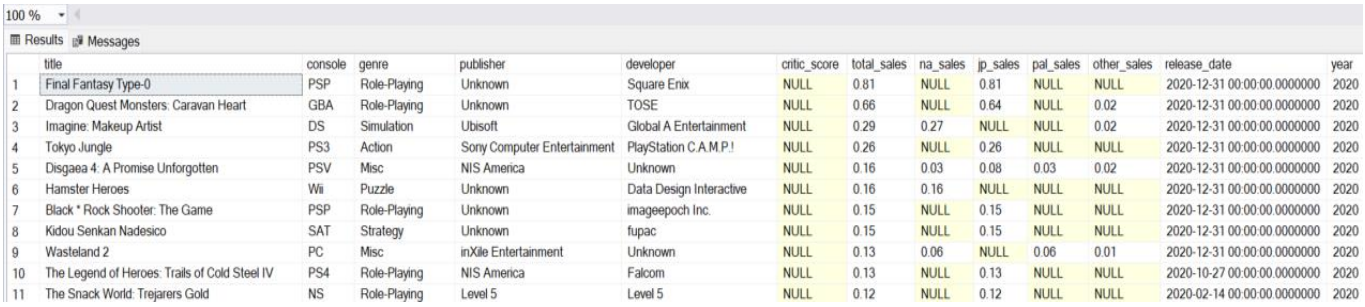
Exploratory Analysis

1. The datasets are first opened in Microsoft Excel and unnecessary columns are deleted.
    a. “game_info”:
        i. We will delete:
            1. “id”
            2. “updated”
            3. “esrb_rating” (as this column is mostly empty)
    b. “game_sales”:
        i. We will delete:

1. “img”
2. “last_update”

2. The data is then imported to Microsoft SQL Server.
    a. Columns of interest are selected individually for analysis within the “game_sales” table.
    b. The columns are then filtered to show titles created within 5 years.
        i. Overall, the results do not demonstrate a sufficient observation number for analysis given the size of both datasets. Therefore, the year scope must be expanded.
    c. In the results below, the data column appears incorrect for the “game_sales” table.
        i. ‘Final Fantasy Type 0’ is a PSP game released in 2011 according to Google.
        ii. Additionally, ‘Tokyo Jungle’ is a PS3 game released in 2012.
    d. Furthermore, the majority of sales columns contain insufficient data for analysis.

```
SELECT title, console, genre, publisher, developer, critic_score,
total_sales, na_sales, jp_sales, pal_sales, other_sales, release_date,
          LEFT(release_date, 4) AS year
       FROM game_sales
       WHERE LEFT(release_date, 4) > 2019
```



100 %

Results | Messages

	title	console	genre	publisher	developer	critic_score	total_sales	na_sales	jp_sales	pal_sales	other_sales	release_date	year
1	Final Fantasy Type-0	PSP	Role-Playing	Unknown	Square Enix	NULL	0.81	NULL	0.81	NULL	NULL	2020-12-31 00:00:00.0000000	2020
2	Dragon Quest Monsters: Caravan Heart	GBA	Role-Playing	Unknown	TOSE	NULL	0.66	NULL	0.64	NULL	0.02	2020-12-31 00:00:00.0000000	2020
3	Imagine: Makeup Artist	DS	Simulation	Ubisoft	Global A Entertainment	NULL	0.29	0.27	NULL	NULL	0.02	2020-12-31 00:00:00.0000000	2020
4	Tokyo Jungle	PS3	Action	Sony Computer Entertainment	PlayStation C.A.M.P.!	NULL	0.26	NULL	0.26	NULL	NULL	2020-12-31 00:00:00.0000000	2020
5	Disgaea 4: A Promise Unforgotten	PSV	Misc	NIS America	Unknown	NULL	0.16	0.03	0.08	0.03	0.02	2020-12-31 00:00:00.0000000	2020
6	Hamster Heroes	Wii	Puzzle	Unknown	Data Design Interactive	NULL	0.16	0.16	NULL	NULL	NULL	2020-12-31 00:00:00.0000000	2020
7	Black * Rock Shooter: The Game	PSP	Role-Playing	Unknown	imageepoch Inc.	NULL	0.15	NULL	0.15	NULL	NULL	2020-12-31 00:00:00.0000000	2020
8	Kidou Senkan Nadesico	SAT	Strategy	Unknown	fupac	NULL	0.15	NULL	0.15	NULL	NULL	2020-12-31 00:00:00.0000000	2020
9	Wasteland 2	PC	Misc	inXile Entertainment	Unknown	NULL	0.13	0.06	NULL	0.06	0.01	2020-12-31 00:00:00.0000000	2020
10	The Legend of Heroes: Trails of Cold Steel IV	PS4	Role-Playing	NIS America	Falcom	NULL	0.13	NULL	0.13	NULL	NULL	2020-10-27 00:00:00.0000000	2020
11	The Snack World: Trejarers Gold	NS	Role-Playing	Level 5	Level 5	NULL	0.12	NULL	0.12	NULL	NULL	2020-02-14 00:00:00.0000000	2020

3. The date column is assessed from “game_sales” by counting the number of rows and grouping by “year.”
    a. The results demonstrate few titles are included after year 2020. Therefore, recent genre and console trends cannot be properly included for analysis.

```
WITH CTE AS (
SELECT title, console, genre, publisher, developer, critic_score,
total_sales, na_sales, jp_sales, pal_sales, other_sales, release_date,
          LEFT(release_date, 4) AS year
       FROM game_sales
)

SELECT DISTINCT year,
          COUNT(*) AS count
       FROM CTE
       GROUP BY year
       ORDER by year DESC
```

	year	count
1	2024	22
2	2023	108
3	2022	168
4	2021	422
5	2020	1450
6	2019	1288
7	2018	1537
8	2017	1557
9	2016	1348
10	2015	1673
11	2014	2891
12	2013	1734
13	2012	1580
14	2011	3383
15	2010	3585
16	2009	4354
17	2008	2923
18	2007	2518
19	2006	2091
20	2005	1809
21	2004	1597

4. The “game_sales” table will be examined more granularly to consider the possibility of working around accurate release date information.
    a. In the query below, the “year” filter scope has been increased to encompass 7th generation console releases—which includes the Microsoft Xbox 360 and Sony PlayStation 3.
    b. Furthermore, target consoles released within 2005 to present day have been selected.
        i. However, the query below cannot be used to select consoles within the ideal release scope—as 'PC' releases will include titles older than 2005 due to the inaccurate release data. Therefore, the release data issue must later be addressed.

```
SELECT title, console, genre, publisher, developer, critic_score,
total_sales, na_sales, jp_sales, pal_sales, other_sales,
          LEFT(release_date, 4) AS year
       FROM game_sales
       WHERE LEFT(release_date, 4) >= 2005
       AND console IN ('Wii', 'WiiU', 'X360', 'XOne', 'XS', 'ZXS',
'PS3', 'PS4', 'PS5', 'iOS', 'PC')
```

	title	console	genre	publisher	developer	critic_score	total_sales	na_sales	jp_sales	pal_sales	other_sales	year
1	Grand Theft Auto V	PS3	Action	Rockstar Games	Rockstar North	9.4	20.32	6.37	0.99	9.85	3.12	2013
2	Grand Theft Auto V	PS4	Action	Rockstar Games	Rockstar North	9.7	19.39	6.06	0.6	9.71	3.02	2014
3	Grand Theft Auto V	X360	Action	Rockstar Games	Rockstar North	NULL	15.86	9.06	0.06	5.33	1.42	2013
4	Call of Duty: Black Ops 3	PS4	Shooter	Activision	Treyarch	8.1	15.09	6.18	0.41	6.05	2.44	2015
5	Call of Duty: Modern Warfare 3	X360	Shooter	Activision	Infinity Ward	8.7	14.82	9.07	0.13	4.29	1.33	2011
6	Call of Duty: Black Ops	X360	Shooter	Activision	Treyarch	8.8	14.74	9.76	0.11	3.73	1.14	2010
7	Red Dead Redemption 2	PS4	Action-Adventure	Rockstar Games	Rockstar Games	9.8	13.94	5.26	0.21	6.21	2.26	2018
8	Call of Duty: Black Ops II	X360	Shooter	Activision	Treyarch	8.4	13.86	8.27	0.07	4.32	1.2	2012
9	Call of Duty: Black Ops II	PS3	Shooter	Activision	Treyarch	8	13.8	4.99	0.65	5.88	2.28	2012
10	Call of Duty: Modern Warfare 2	X360	Shooter	Activision	Infinity Ward	9.5	13.53	8.54	0.08	3.63	1.28	2009
11	Call of Duty: WWII	PS4	Shooter	Activision	Sledgehammer Games	8.1	13.4	4.67	0.4	6.21	2.12	2017

5. The “game_info” table is assessed for results.
    a. This table does not create new rows for individual releases by console.
        i. However, the “game_sales” and “game_info” tables will be joined together—which will solve this issue.
    b. The data range is increased to match the “game_sales” table, where the most information is derived.
    c. Overall, the data column in the “game_info” table appears more accurate.
        i. Upon cross-checking 10 title release dates using Google, the date column in this dataset appears more accurate.
        ii. Therefore, a join is performed to retain the date column from the “game_info” table.

```
SELECT "name", suggestions_count, platforms, developers, genres,
publishers,
   LEFT(released, 4) AS year
FROM game_info
WHERE LEFT(released, 4) >= 2005
AND "platforms" LIKE '%PlayStation 5%' OR "platforms" LIKE '%Xbox
Series%'
```

	name	suggestions_count	platforms	developers	genres	publishers	year
1	Maximum Override	155	Xbox One||PlayStation 5||PC||Xbox 360||iOS||Andr...	Alientrap||y8.com	Action||Shooter||Simulation||Casual||Indie	Alientrap||y8.com	2017
2	Hogwarts Legacy	621	PC||Xbox One||PlayStation 4||Xbox Series S/X||Pla...	Avalanche Software||Portkey Games	Adventure||RPG	Warner Bros. Interactive||Wizarding World	NULL
3	Warhammer: Chaosbane	538	PC||Xbox Series S/X||PlayStation 5||Xbox One||Pla...	Eko Software	Action||Adventure||RPG	Bigben Interactive||Nacon	2019
4	Dragon Age 4: The Dread Wolf Rises	42	PlayStation 5||Xbox Series S/X||PC	BioWare	RPG	Electronic Arts	NULL
5	Paradise Lost	633	PC||Xbox Series S/X||PlayStation 5	PolyAmorous	Adventure||Indie	All in! Games	NULL
6	Ori and the Will of the Wisps	465	PC||Nintendo Switch||Xbox Series S/X||Xbox One	Moon Studios	Action||Adventure||Platformer	Microsoft Studios||Xbox Game Studios	2020
7	Oddworld: Soulstorm	232	PC||PlayStation 4||PlayStation 5	Oddworld Inhabitants	Action||Adventure||Indie||Platformer	Oddworld Inhabitants	2021
8	Haven	256	PlayStation 4||Xbox Series S/X||Xbox One||PlaySta...	The Game Bakers	Adventure||RPG	The Game Bakers	2020
9	Recompile	367	Xbox Series S/X||PlayStation 5||PC	Phigames	Action||Adventure||Indie	Dear Villagers	NULL
10	Vampire: The Masquerade - Bloodlines 2	582	PlayStation 5||Xbox Series S/X||PC||PlayStation 4||...	Hardsuit Labs	Action||RPG	Paradox Interactive	2021
11	Destiny 2	675	PlayStation 5||Web||Xbox Series S/X||PC||Xbox On...	Vicarious Visions	Action||Shooter||Massively Multiplayer	Activision Blizzard||Bungie	2017
12	Watch Dogs Legion	558	Xbox Series S/X||PlayStation 5||Xbox One||PlaySta...	Ubisoft!!Ubisoft Toronto	Action!!Shooter!!Adventure	Ubisoft Entertainment	2020

Data Cleaning

1. The “released” column from the “game_info” table is changed to “date” data type.

```
ALTER TABLE game_info ADD released_converted date;

Update game_info
SET released_converted = CONVERT(date, released)

SELECT "name", suggestions_count, platforms, developers, genres, publishers,
       YEAR(released_converted) AS year_2
       FROM game_info
       WHERE YEAR(released_converted) >= 2005
```

2. Duplicates are found and dropped across both datasets.
   a. Duplicates are only dropped under logical circumstances. In this case, duplicate observations are dropped as repeat release observations for the same console are unnecessary.
   b. Therefore, duplicates within “game_sales” are located.

```
WITH CTE AS(
SELECT *,
       ROW_NUMBER() OVER (
       PARTITION BY title,
                    console,
                          publisher,
                          developer
                ORDER BY title) AS duplicate_finder
FROM game_sales)

SELECT *
FROM CTE
WHERE duplicate_finder > 1
```

   c. Duplicates within “game_sales” are then deleted.

```
WITH CTE AS(
SELECT *,
       ROW_NUMBER() OVER (
       PARTITION BY title,
                        console,
                              publisher,
                              developer
                ORDER BY title) AS duplicate_finder
FROM game_sales)

DELETE
FROM CTE
WHERE duplicate_finder > 1
```

d. Duplicates within the “game_info” dataset are located.

```
WITH CTE AS (
SELECT *,
          ROW_NUMBER() OVER(
          PARTITION BY "name",
                                developers
                     ORDER BY id) AS duplicate_finder
       FROM game_info)

SELECT *
FROM CTE
WHERE duplicate_finder > 1
```

e. Duplicates within the “game_info” dataset are deleted.

```
WITH CTE AS (
SELECT *,
          ROW_NUMBER() OVER(
          PARTITION BY "name",
                                developers
                     ORDER BY id) AS duplicate_finder
       FROM game_info)

DELETE
FROM CTE
WHERE duplicate_finder > 1
```

3. An inner join is performed across the “game_sales” and “game_info” tables.
    a. A join is created using video game titles to create a join upon.
        i. Additionally, an inner join is used to drop titles without matches from both tables.
        ii. A WHERE clause is used to filter for rows with at least 1 column with regional sales; observations with entirely absent regional sales columns cannot be used for analysis.
        iii. The “platforms” column from the “game_info” table appears more accurate than the “console” column from the “game_sales” table.
            1. ‘Sudden Strike 4’ was released for macOS||iOS||Linux||PC||PlayStation 4
            2. ‘Cities:Skylines’ was released for Linux||macOS||PC||Nintendo Switch||Xbox One||PlayStation 4
                a. However, the ‘platforms’ column cannot be used from “game_info” as the “game_sales” table contains the most important sales data.

```
WITH CTE AS(
SELECT title, console, genre, publisher, developer,
critic_score, total_sales, na_sales, jp_sales,
pal_sales, other_sales,
          YEAR(release_date_converted) AS year
       FROM game_sales
       WHERE YEAR(release_date_converted) >= 2005),

CTE_2 AS(SELECT "name", suggestions_count, platforms,
developers, genres, publishers,
       YEAR(released_converted) AS year_accurate
       FROM game_info
       WHERE YEAR(released_converted) >= 2005)

SELECT CTE.title, genre, console, publisher, developer,
critic_score, total_sales, na_sales, jp_sales,
pal_sales, other_sales,
          CTE_2.platforms, year_accurate
       FROM CTE
       INNER JOIN CTE_2
       ON CTE.title = CTE_2.name
       WHERE CTE.total_sales IS NOT NULL
```

title	genre	console	publisher	developer	critic_score	total_sales	na_sales	jp_sales	pal_sales	other_sales	platforms	year_accurate
Sudden Strike 4	Strategy	PS4	Kalypso Media	Kite Games	NULL	0.1	0.05	0.01	0.03	0.02	macOS||iOS||Linux||PC||PlayStation 4	2017
Sudden Strike 4	Strategy	XOne	Kalypso Media	Kite Games	NULL	0.02	0.02	NULL	NULL	0	macOS||iOS||Linux||PC||PlayStation 4	2017
Hitman	Action	PS4	Square Enix	IO Interactive	NULL	0.78	0.24	0.06	0.36	0.11	Linux||macOS||PC||Xbox One||PlayStation 4	2017
Hitman	Action	XOne	Square Enix	IO Interactive	NULL	0.31	0.2	NULL	0.08	0.03	Linux||macOS||PC||Xbox One||PlayStation 4	2017
Tyranny	Role-Playing	PC	Paradox Interactive	Obsidian Entertainment	NULL	0.02	NULL	NULL	0.02	0	macOS||Linux||PC	2016
Project: Snowblind	Action	PS2	Eidos Interactive	Unknown	NULL	0.21	0.1	NULL	0.08	0.03	PC	2005
Project: Snowblind	Action	XB	Eidos Interactive	Unknown	NULL	0.11	0.08	NULL	0.02	0	PC	2005
Project: Snowblind	Shooter	PC	Eidos Interactive	Crystal Dynamics	NULL	0	NULL	NULL	0	0	PC	2005
Order of War	Strategy	PC	Square Enix	Wargaming.net	NULL	0.02	0	NULL	0.01	0	PC	2009
Watch Dogs 2	Action	PS4	Ubisoft	Ubisoft	NULL	3.36	0.98	0.12	1.74	0.52	PC||PlayStation 4||Xbox One	2016
Watch Dogs 2	Action	XOne	Ubisoft	Ubisoft	NULL	1.35	0.71	NULL	0.53	0.12	PC||PlayStation 4||Xbox One	2016
Watch Dogs 2	Action	PC	Ubisoft	Ubisoft	NULL	0	NULL	NULL	0	0	PC||PlayStation 4||Xbox One	2016
Cities: Skylines	Simulation	PS4	Koch Media	Colossal Order	NULL	0.3	0.15	0.04	0.07	0.05	Linux||macOS||PC||Nintendo Switch||Xbox One||Play...	2015
Cities: Skylines	Simulation	XOne	Koch Media	Colossal Order	NULL	0.18	0.13	NULL	0.03	0.02	Linux||macOS||PC||Nintendo Switch||Xbox One||Play...	2015
Owlboy	Platform	PS4	Soedesco	D-Pad Studios	NULL	0.05	0.04	NULL	NULL	0.01	PlayStation 4||Nintendo Switch||Linux||PC||Xbox One	2016
Owlboy	Platform	NS	Soedesco	D-Pad Studios	9	0.05	0.03	NULL	0.01	0.01	PlayStation 4||Nintendo Switch||Linux||PC||Xbox One	2016
Batman: Return to Arkham	Action	PS4	Warner Bros. Interactive Entertainment	Rocksteady Studios	NULL	0.64	0.15	0.01	0.39	0.1	Xbox One||PlayStation 4	2016
Batman: Return to Arkham	Action	XOne	Warner Bros. Interactive Entertainment	Rocksteady Studios	NULL	0.2	0.12	NULL	0.06	0.02	Xbox One||PlayStation 4	2016
Portal Knights	Role-Playing	PS4	505 Games	Keen Games	NULL	0.12	0.02	0.03	0.05	0.01	Android||iOS||PC||Xbox One||PlayStation 4||Nintend...	2016
Portal Knights	Role-Playing	NS	505 Games	Keen Games	NULL	0.06	0.02	0.01	0.02	0	Android||iOS||PC||Xbox One||PlayStation 4||Nintend...	2016
Portal Knights	Role-Playing	XOne	505 Games	Keen Games	NULL	0.02	0.02	NULL	NULL	0	Android||iOS||PC||Xbox One||PlayStation 4||Nintend...	2016
RPG Maker MV	Misc	PS4	NIS America	Kadokawa Games	NULL	0.01	NULL	0.01	NULL	NULL	PlayStation 4||Linux||PC||macOS	2015

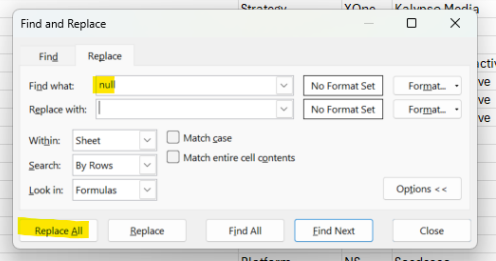
4. Missing data must now be addressed.
    a. Many missing values exist for regional sales data that require imputation.
    b. Statistically modeling datasets with missing values is not possible without addressing the issue.
        i. Overall, the data provided is difficult to statistically model as the independent variables—which include title, console, genre and publisher—are all nominal and categorical. Therefore, dummy variables would be necessary to assign binary values to nominal data. In this case, a myriad of nominal variables exist. Therefore, statistical modelling is not easily feasible.
    c. Within SQL, imputation is primarily performed using mean or median of the entire dataset.
        i. However, this will lead to more skewed results than using different imputation methods.
        ii. Therefore, the data is imputed using R Studio.

```
WITH CTE AS(
SELECT title, console, genre, publisher, developer, critic_score,
total_sales, na_sales, jp_sales, pal_sales, other_sales,
          YEAR(release_date_converted) AS year
       FROM game_sales
       WHERE YEAR(release_date_converted) >= 2005),

CTE_2 AS(SELECT "name", suggestions_count, platforms, developers,
genres, publishers,
       YEAR(released_converted) AS year_accurate
       FROM game_info
       WHERE YEAR(released_converted) >= 2005)

SELECT CTE.title, genre, console, publisher, developer,
critic_score, total_sales, na_sales, jp_sales, pal_sales,
other_sales,
          CTE_2.platforms, year_accurate
       FROM CTE
       INNER JOIN CTE_2
       ON CTE.title = CTE_2.name
       WHERE CTE.total_sales IS NOT NULL
```

5. Initially, the data is exported to Microsoft Excel.
    a. NULL values are replaced with empty values in Microsoft Excel. Otherwise, they will not be detected as empty within R Studio.



title	genre	console	publisher	developer
Sudden Strike 4	Strategy	PS4	Kalypso Media	Kite Games
Sudden Strike 4	Strategy	XOne	Kalypso Media	Kite Games
Hitman	Action	PS4		IO Interactive
Hitman	Action	XOne		IO Interactive
Tyranny	Role-Playing	PC	ctive	Obsidian Entertainment
Project: Snowblind	Action	PS2	ve	Unknown
Project: Snowblind	Action	XBox	ve	Unknown
Project: Snowblind	Action	PC	ve	Crystal Dynamics
Order of War	Strategy	PC		Wargaming.net
Watch Dogs 2	Action	PS4		Ubisoft
Watch Dogs 2	Action	XOne		Ubisoft
Watch Dogs 2	Action	PC		Ubisoft
Cities: Skylines	Simulation	PS4		Colossal Order
Cities: Skylines	Simulation	XOne		Colossal Order
Owlboy	Platform	PC		D-Pad Studios
Owlboy	Platform	NS	Soedesco	D-Pad Studios
Batman: Return to Arkham	Action	PS4	Warner Bros. Interactive Entertainment	Rocksteady Studios
Batman: Return to Arkham	Action	XOne	Warner Bros. Interactive Entertainment	Rocksteady Studios
Portal Knights	Role-Playing	PS4	505 Games	Keen Games
Portal Knights	Role-Playing	NS	505 Games	Keen Games
Portal Knights	Role-Playing	XOne	505 Games	Keen Games

6. Within R Studio:
    a. The necessary packages are run.

```
knitr::opts_chunk$set(echo = TRUE)
library(VIM)
library(mice)
library(dplyr)
library(datarium)
game_data <- read.csv("C:/Users/emers/Desktop/Video Game
Analysis/Video_Game_Analysis_Cleaned.csv")
```

    b. NA values are located and counted.

```
sapply(game_data, function(x) sum(is.na(x)))
```

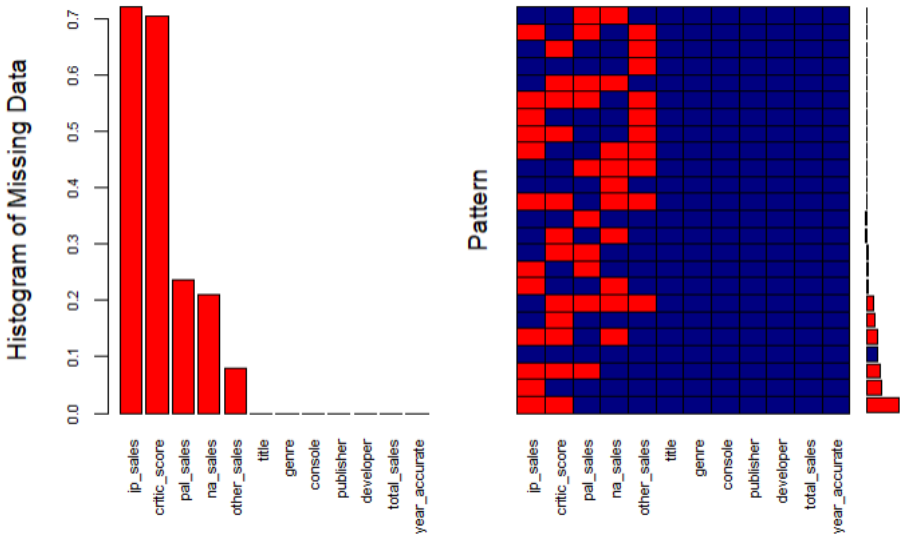
```
      title       genre     console   publisher   developer critic_score total_sales
          0           0           0           0           0        5651           0
   na_sales    jp_sales   pal_sales other_sales year_accurate
       1688        5787        1902         641           0
```

    c. Percentages of NA values for each column are assessed.

```
sapply(game_data, function(x) mean(is.na(x)) * 100)
```

```
      title       genre     console   publisher   developer critic_score total_sales
   0.000000    0.000000    0.000000    0.000000    0.000000   70.294813    0.000000
   na_sales    jp_sales   pal_sales other_sales year_accurate
  20.997637   71.986565   23.659659    7.973629    0.000000
```

    d. A visualization to understand the proportion of missing data is created.
        i. Overall, the “jp_sales” and “critic_score” columns are both primarily missing.
        ii. Imputing data should not affect the mean and data distribution; it is always better to add more data if possible.
            1. The “critic_score” column has a high percentage of missing values. Therefore, it will be dropped.
        iii. Generally, imputation performs most effectively if missing values are <= 5% in a column.
        iv. Additionally, imputation should not be used if the data is more than 20% missing.
            1. A large percentage of regional sales data is missing. Additionally, 70% of the “jp_sales” column is absent. Therefore, an exception will be made due to the lack of usable data. Regional sales will then be recombined to create a new total sales column.

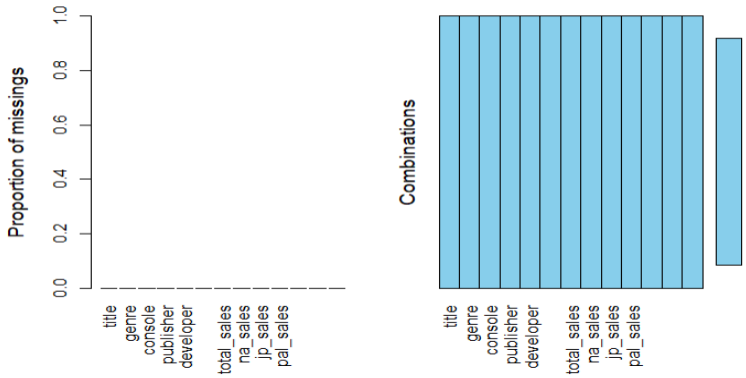


e. Missing data is imputed using the kNN() function.
    i. The K Nearest-Neighbor method is utilized, as this method is the standard for data imputation due to its accuracy and efficacy.

```
game_data_knn <- kNN(game_data, variable = c("jp_sales", "critic_score",
"pal_sales", "na_sales", "other_sales"), k = 6)
```

f. The imputation results are evaluated using an aggregation plot to ensure success.

```
aggr(game_data_knn)
```



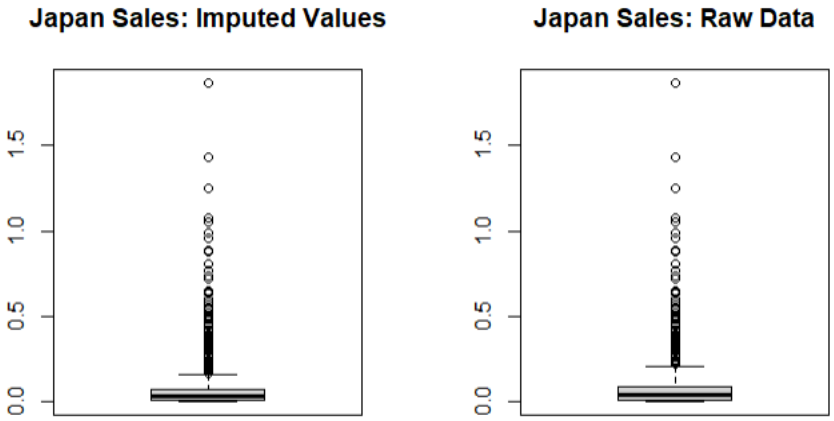
g. NA values are reassessed using the imputed data to ensure absence of missing values.

```
sapply(game_data_knn, function(x) sum(is.na(x)))
```

```
        title         genre       console     publisher     developer  critic_score
            0             0             0             0             0             0
 total_sales      na_sales      jp_sales     pal_sales   other_sales  year_accurate
            0             0             0             0             0             0
jp_sales_imp critic_score_imp pal_sales_imp na_sales_imp other_sales_imp
            0             0             0             0             0
```

h. A boxplot is created to visualize possible distribution differences between the original and imputed data.
    i. Imputed data values should maintain the same distribution as the raw data.

```
par(mfrow=c(1,2))
boxplot(game_data_knn$jp_sales, main = "Japan Sales: Imputed Values")
boxplot(game_data$jp_sales, main = "Japan Sales: Raw Data")
```



i. A t-test is used to compare sample means between the original and imputed data.
    i. The results demonstrate a low p-value indicative of no association between the “jp_sales” variables across both datasets.

```
t.test(game_data_knn$jp_sales, game_data$jp_sales)
```

```
        Welch Two Sample t-test

data:  game_data_knn$jp_sales and game_data$jp_sales
t = -6.6774, df = 2988.4, p-value = 2.889e-11
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.02327149 -0.01270676
sample estimates:
 mean of x  mean of y
0.05786789 0.07585702
```

j. A new column is created containing the combination of values for each regional sales column. This column represents the new “total sales” column.

```
game_data_final <- mutate(game_data_comp, total_sales_cleaned = na_sales +
jp_sales + pal_sales + other_sales
```

critic_score <dbl>	total_sales <dbl>	na_sales <dbl>	jp_sales <dbl>	pal_sales <dbl>	other_sales <dbl>	year_accurate <int>	total_sales_cleaned <dbl>
7.0	0.10	0.05	0.01	0.03	0.02	2017	0.11
9.0	0.02	0.02	0.02	0.02	0.00	2017	0.06
7.0	0.78	0.24	0.06	0.36	0.11	2017	0.77
6.2	0.31	0.20	0.01	0.08	0.03	2017	0.32
9.2	0.02	0.02	0.02	0.02	0.00	2016	0.06
6.2	0.21	0.10	0.00	0.08	0.03	2005	0.21
8.7	0.11	0.08	0.06	0.02	0.00	2005	0.16
4.7	0.00	0.00	0.01	0.00	0.00	2005	0.01
8.0	0.02	0.00	0.02	0.01	0.00	2009	0.03
9.5	3.36	0.98	0.12	1.74	0.52	2016	3.36

7. Before the data is reuploaded to Microsoft SQL Server, the sales columns are converted to actual sales figures.
    a. The sales-related columns are converted to millions sold for intuitive reporting.
    b. Additionally, the “critic_score” and original “total_sales” columns are dropped.

8. The imputed data is reuploaded to Microsoft SQL Server for analysis.

```
SELECT *
FROM game_data_imputed
```

	id	title	genre	console	publisher	developer	na_sales	jp_sales	pal_sales	other_sales	year_accurate	total_sales_cleaned
1	1	Sudden Strike 4	Strategy	PS4	Kalypso Media	Kite Games	50000	10000	30000	20000	2017	110000
2	2	Sudden Strike 4	Strategy	XOne	Kalypso Media	Kite Games	20000	20000	20000	0	2017	60000
3	3	Hitman	Action	PS4	Square Enix	IO Interactive	240000	60000	360000	110000	2017	770000
4	4	Hitman	Action	XOne	Square Enix	IO Interactive	200000	10000	80000	30000	2017	320000
5	5	Tyranny	Role-Playing	PC	Paradox Interactive	Obsidian Entertainment	20000	20000	20000	0	2016	60000
6	6	Project: Snowblind	Action	PS2	Eidos Interactive	Unknown	100000	0	80000	30000	2005	210000
7	7	Project: Snowblind	Action	XB	Eidos Interactive	Unknown	80000	60000	20000	0	2005	160000

Critical Analysis

1. The most ideal genre to create is assessed using the cleaned and imputed dataset.
    a. Total sales by genre are assessed across the dataset.
        i. Overall, action, sports and shooter games are the most popular.

```
WITH CTE AS (
SELECT genre,
          SUM(total_sales_cleaned) AS total_sales_by_genre,
          FORMAT(SUM(total_sales_cleaned), 'C') AS
total_sales_by_genre_$
       FROM game_data_imputed
       GROUP BY genre)

SELECT genre,
       total_sales_by_genre_$
  FROM CTE
  ORDER BY total_sales_by_genre DESC
```

	genre	total_sales_by_genre_$
1	Action	$705,630,000.00
2	Sports	$672,930,000.00
3	Shooter	$652,550,000.00
4	Misc	$323,030,000.00
5	Role-Playing	$274,780,000.00
6	Racing	$230,510,000.00
7	Adventure	$194,490,000.00
8	Simulation	$173,260,000.00
9	Platform	$158,100,000.00
10	Fighting	$140,540,000.00
11	Action-Adventure	$97,440,000.00
12	Puzzle	$59,030,000.00
13	Strategy	$55,280,000.00
14	Music	$50,380,000.00

2. Highest-selling individual releases are then assessed.
    a. Action and shooter games are among the most popular individual releases.
        i. Interestingly, sports games appear less popular as individual releases.
            1. This must be considered, as the goal is to create one single title.
    b. Additionally, the Sony PlayStation 3 and Microsoft Xbox 360 console releases are the most popular.
        i. This indicates 7th generation consoles—which include the PlayStation 3 and Xbox 360—may be more popular than their 8th generation counterparts—which include the PlayStation 4 and Xbox One.

```
WITH CTE AS (
SELECT title, genre, console, publisher, developer, year_accurate,
total_sales_cleaned,
          FORMAT(total_sales_cleaned, 'C') AS total_sales_$
       FROM game_data_imputed
       )

SELECT title, genre, console, publisher, developer, year_accurate,
total_sales_$
       FROM CTE
       ORDER BY total_sales_cleaned DES
```

Results Messages

	title	genre	console	publisher	developer	year_accurate	total_sales_$
1	Grand Theft Auto V	Action	PS3	Rockstar Games	Rockstar North	2013	$20,330,000.00
2	Grand Theft Auto V	Action	PS4	Rockstar Games	Rockstar North	2013	$19,390,000.00
3	Grand Theft Auto V	Action	X360	Rockstar Games	Rockstar North	2013	$15,870,000.00
4	Call of Duty: Modern Warfare 3	Shooter	X360	Activision	Infinity Ward	2011	$14,820,000.00
5	Call of Duty: Black Ops	Shooter	X360	Activision	Treyarch	2010	$14,740,000.00
6	Red Dead Redemption 2	Action-Adventure	PS4	Rockstar Games	Rockstar Games	2018	$13,940,000.00
7	Call of Duty: Black Ops II	Shooter	X360	Activision	Treyarch	2012	$13,860,000.00
8	Call of Duty: Black Ops II	Shooter	PS3	Activision	Treyarch	2012	$13,800,000.00
9	Call of Duty: Modern Warfare 2	Shooter	X360	Activision	Infinity Ward	2009	$13,530,000.00
10	Call of Duty: WWII	Shooter	PS4	Activision	Sledgehammer Games	2017	$13,400,000.00
11	Call of Duty: Modern Warfare 3	Shooter	PS3	Activision	Infinity Ward	2011	$13,350,000.00
12	Call of Duty: Black Ops	Shooter	PS3	Activision	Treyarch	2010	$12,670,000.00
13	FIFA 18	Sports	PS4	EA Sports	EA Vancouver	2017	$11,790,000.00
14	Grand Theft Auto IV	Action	X360	Rockstar Games	Rockstar North	2008	$11,090,000.00
15	FIFA 17	Sports	PS4	Electronic Arts	EA Canada	2016	$10,940,000.00
16	Call of Duty: Modern Warfare 2	Shooter	PS3	Activision	Infinity Ward	2009	$10,620,000.00
17	Grand Theft Auto IV	Action	PS3	Rockstar Games	Rockstar North	2008	$10,580,000.00

3. A query is created to assess genre count across releases for developers with the highest percentage of sales.
    a. Within the query, duplicates are addressed—which currently exist as titles separated into individual rows based on console.

```
SELECT DISTINCT title,
          developer,
          genre
       FROM game_data_imputed
       WHERE developer = 'Rockstar North'
```

title	developer	genre
Grand Theft Auto IV	Rockstar North	Action
Grand Theft Auto V	Rockstar North	Action
Grand Theft Auto: Episodes from Liberty City	Rockstar North	Action
Grand Theft Auto: Episodes from Liberty City	Rockstar North	Adventure
Grand Theft Auto: The Trilogy	Rockstar North	Adventure

    b. Therefore, a CTE is created to contain only first instances of game titles using the query below.

```
WITH CTE AS(
SELECT *,
       ROW_NUMBER() OVER (
       PARTITION BY title
              ORDER BY title) AS duplicate_finder
FROM game_data_imputed
)

SELECT *
FROM CTE
WHERE duplicate_finder = 1
```

id	title	genre	console	publisher	developer	na_sales	jp_sales	pal_sales	other_sales	year_accurate	total_sales_cleaned	column13	column14	duplicate_finder
6621	.hack//G.U. Last Recode	Role-Playing	PS4	Namco Bandai Games	CyberConnect2	100000	80000	40000	30000	2017	250000	NULL	NULL	1
4181	.hack//Link	Role-Playing	PSP	Namco Bandai	CyberConnect2	140000	140000	0	10000	2010	290000	NULL	NULL	1
7179	007 Legends	Shooter	PS3	Activision	Eurocom	110000	60000	170000	60000	2012	400000	NULL	NULL	1
2355	1 vs. 100	Misc	DS	DSI Games	ECI	80000	0	0	10000	2008	90000	NULL	NULL	1
4204	10 Minute Solution	Sports	Wii	Activision	Anchor Bay Entertainment	60000	20000	10000	10000	2010	100000	NULL	NULL	1
3038	100 Classic Books	Misc	DS	Nintendo	Genius Sonority Inc.	120000	550000	520000	20000	2010	1210000	NULL	NULL	1
3251	100 Classic Games	Misc	DS	Rondomedia	Easy Interactive	40000	40000	30000	0	2012	110000	NULL	NULL	1

    c. A query is then created to count the number of genres by developer and join the tables together using CTEs.
        i. According to the query below, ‘EA%’ has the highest number of total sales by developer—and most frequently produces sports titles.
            1. The table below indicates developers with the highest percentage of sales most frequently create sports or action games.

```
WITH CTE AS (
SELECT *,
       ROW_NUMBER() OVER (
       PARTITION BY title
              ORDER BY title) AS duplicate_finder
FROM game_data_imputed
),
```

```
CTE_2 AS (
SELECT *
FROM CTE
WHERE duplicate_finder = 1
),

CTE_3 AS (
SELECT developer,
          genre,
       COUNT(genre) AS genre_count
       FROM CTE_2
       GROUP BY developer, genre
),

CTE_4 AS (
SELECT developer,
       CAST(100 * SUM(total_sales_cleaned)/(SELECT
SUM(total_sales_cleaned) FROM game_data_imputed) AS
DECIMAL(7,2)) AS percentage_of_total_sales_by_developer
       FROM game_data_imputed
       GROUP BY developer
)

SELECT
             CTE_3.developer,
             CTE_3.genre,
             CTE_4.percentage_of_total_sales_by_developer,
          CTE_3.genre_count
       FROM CTE_4
       INNER JOIN CTE_3
       ON CTE_3.developer = CTE_4.developer
       WHERE genre_count >= 10
       ORDER BY CTE_4.percentage_of_total_sales_by_developer
DESC, genre_count DESC
```

	developer	genre	percentage_of_total_sales_by_developer	genre_count
1	EA Canada	Sports	4.38	45
2	EA Tiburon	Sports	3.52	36
3	Ubisoft Montreal	Action	2.43	10
4	Traveller's Tales	Action	2.40	14
5	Visual Concepts	Sports	2.17	23
6	Capcom	Action	1.31	16
7	Capcom	Fighting	1.31	10
8	Konami	Sports	1.17	12
9	Ubisoft	Misc	1.12	18
10	EA Redwood Shores	Simulation	1.02	10

4. To assess consoles to be included for the release, the percentage of total sales by console are evaluated.
    a. Overall, the Microsoft Xbox 360 and Sony PlayStation 3 are the most popular.
        i. Console releases generate a far greater percentage of sales than PC titles.

ii. Additionally, ninth generation consoles—which include the Microsoft Xbox Series X/S and Sony PlayStation 5 consoles—were only released in 2020. As discussed, the latest usable data from the "game_sales" table is from 2020. Therefore, ninth generation console sales are not possible to effectively analyze.

iii. The Sony PlayStation 4 and Microsoft Xbox One sold far less titles than their predecessors.

1. Therefore, developing a PC-only title may generate the highest return on investment given the generational decrease in console sales.

```
SELECT console,
       CAST(100 * SUM(total_sales_cleaned)/(SELECT
SUM(total_sales_cleaned) FROM game_data_imputed) AS DECIMAL(7,2)) AS
percentage_of_total_sales_by_console
       FROM game_data_imputed
       GROUP BY console
       ORDER BY CAST(100 * SUM(total_sales_cleaned)/(SELECT
SUM(total_sales_cleaned) FROM game_data_imputed) AS DECIMAL(7,2))
DESC
```

	console	percentage_of_total_sales_by_console
1	X360	19.74
2	PS3	18.99
3	PS4	11.96
4	Wii	10.28
5	DS	9.25
6	PS2	6.93
7	XOne	6.16
8	PSP	4.96
9	PC	3.56
10	3DS	1.74
11	XB	1.57
12	PSV	1.44
13	NS	0.97
14	GC	0.93

5. Next, the percentage of total sales by console and developer are assessed.

a. 'Infinity Ward' and 'Treyarch' are well-known developers of shooter games; 'EA%' generally produces sports titles.

i. Therefore, the query further demonstrates the popularity of these titles.

ii. The query also indicates the popularity of sports and shooter games for 7th generation consoles.

```
SELECT developer,
          console,
       CAST(100 * SUM(total_sales_cleaned)/(SELECT
SUM(total_sales_cleaned) FROM game_data_imputed) AS DECIMAL(7,2)) AS
percentage_of_total_sales_by_developer
       FROM game_data_imputed
       GROUP BY console, developer
       ORDER BY CAST(100 * SUM(total_sales_cleaned)/(SELECT
SUM(total_sales_cleaned) FROM game_data_imputed) AS DECIMAL(7,2))
DESC
```

	developer	console	percentage_of_total_sales_by_developer
1	Infinity Ward	X360	1.32
2	EA Canada	PS3	1.10
3	Infinity Ward	PS3	1.07
4	Treyarch	X360	1.04
5	EA Tiburon	X360	0.96
6	Treyarch	PS3	0.90
7	Rockstar North	PS3	0.87
8	EA Canada	X360	0.86
9	EA Tiburon	PS3	0.78
10	Rockstar North	X360	0.78

6. A query is created to identify trends by visualizing most sold releases for each year included in the dataset.
    a. Overall, the query demonstrates the popularity of shooter games has decreased over time.
    b. Additionally, sports and action/action-adventure titles are likely to sell the most copies.

```
WITH CTE_1 AS (
SELECT developer AS developer,
       genre,
          title,
          year_accurate,
          CAST(100 * SUM(total_sales_cleaned)/(SELECT
SUM(total_sales_cleaned) FROM game_data_imputed) AS DECIMAL(7,2)) AS
percentage_of_total_sales
       FROM game_data_imputed
       GROUP BY year_accurate, developer, genre, title),

CTE_2 AS (SELECT year_accurate,
          genre,
          title,
          developer,
          percentage_of_total_sales,
          RANK() OVER (PARTITION BY year_accurate ORDER BY
percentage_of_total_sales DESC) AS year_rank
FROM CTE_1)

SELECT year_accurate,
          genre,
          title,
          developer
       FROM CTE_2
       WHERE year_rank = 1
       ORDER BY year_accurate DESC
```

	year_accurate	genre	title	developer
1	2020	Sports	Skate	EA Black Box
2	2019	Adventure	Up	THQ
3	2018	Action-Adventure	Red Dead Redemption 2	Rockstar Games
4	2017	Shooter	Call of Duty: WWII	Sledgehammer Games
5	2016	Sports	FIFA 17	EA Canada
6	2015	Role-Playing	Fallout 4	Bethesda Game Studios
7	2014	Sports	FIFA 15	EA Canada
8	2013	Action	Grand Theft Auto V	Rockstar North
9	2012	Shooter	Call of Duty: Black Ops II	Treyarch
10	2011	Shooter	Call of Duty: Modern Warfare 3	Infinity Ward
11	2010	Shooter	Call of Duty: Black Ops	Treyarch
12	2009	Shooter	Call of Duty: Modern Warfare 2	Infinity Ward
13	2008	Action	Grand Theft Auto IV	Rockstar North
14	2007	Shooter	Call of Duty 4: Modern Warfare	Infinity Ward
15	2006	Misc	Guitar Hero II	Harmonix Music Systems
16	2005	Action	Grand Theft Auto: Liberty City Stories	Rockstar Leeds

7. A query is created to visualize the top 3 highest selling genres by year.
    a. The data demonstrates sports and action games are among the highest consistently performing genres across the years.
    b. 2020 represents the latest year available. Therefore, sports games are likely the most consistently high performing titles.
        i. Other genres such as ‘Role-Playing’ and ‘Shooter’ titles are likely to trend.

```
WITH CTE_1 AS (
SELECT genre,
          year_accurate,
          CAST(100 * SUM(total_sales_cleaned)/(SELECT
SUM(total_sales_cleaned) FROM game_data_imputed) AS DECIMAL(7,2)) AS
percentage_of_total_sales
       FROM game_data_imputed
       GROUP BY year_accurate, genre
),

CTE_2 AS (SELECT year_accurate,
          genre,
          percentage_of_total_sales,
          RANK() OVER (PARTITION BY year_accurate ORDER BY
percentage_of_total_sales DESC) AS year_rank
FROM CTE_1
)

SELECT year_accurate,
          genre,
          year_rank
FROM CTE_2
WHERE percentage_of_total_sales != 0.00
AND year_rank IN (1, 2, 3)
ORDER BY year_accurate DESC
```

	year_accurate	genre	year_rank
1	2020	Sports	1
2	2020	Puzzle	2
3	2020	Role-Playing	3
4	2019	Adventure	1
5	2019	Racing	2
6	2019	Role-Playing	3
7	2018	Action-Adventure	1
8	2018	Sports	2
9	2018	Action	3
10	2017	Shooter	1
11	2017	Sports	2
12	2017	Action	3
13	2016	Shooter	1
14	2016	Sports	2
15	2016	Action	3
16	2015	Action	1
17	2015	Sports	2
18	2015	Role-Playing	3
19	2014	Shooter	1
20	2014	Sports	2
21	2014	Role-Playing	3
22	2013	Action	1

Conclusion

Overall, only genre and console analysis could be properly conducted. The analysis demonstrates console releases generate the highest number of sales. Additionally, sports and action titles are among the most predictably popular; the most popular developers typically create these genres. Furthermore, shooter and role-playing-game titles trend more and are less predictable when estimating potential sales.
The decrease in console sales from 7th to 8th generation must also be considered; a far larger percentage of sales (according to the data) consist of 7th generation console sales.

Based on findings of this analysis only, a sports video game for PC and current-generation consoles represents the safest title to create. However, further research must be conducted on the finance associated with video game creation.
Ultimately, the hypothesis that indie survival PC games are the most popular relative to cost cannot be proven or disproven. Therefore, further research must be conducted to answer the questions in blue.