
SQL Final Project
Data Research Analyst

Student: Rina Rafalski

Mentor: Shiran Alon

Q1. Without looking at the following tasks, what kind of business questions do you have in mind after exploring the dataset? Provide at least 2 significant examples verbally and elaborate how you would measure it.

A1.

```
/* Business Question 1:

   What are Popular genres and platforms which have more sales over the years?
-----------------------------------------------------------------------------
   It could be measured as following:
*/

-- Top 5 Genres that have maximum Sales over the years
SELECT TOP 5
       Genre
       ,ROUND(SUM(Global_Sales),1) AS TotalSales
FROM
       video_games
GROUP BY Genre
ORDER BY TotalSales DESC

/* RESULT --------------------------------------------------------------------
Genre                       TotalSales
Action                      1744.2
Sports                      1331.3
Shooter                     1052.4
Role-Playing                934.6
Platform                    827.8
*/

-- Top 5 Platforms that have maximum Sales over the years
SELECT TOP 5
       Platform
       ,ROUND(SUM(Global_Sales),1) AS TotalSales
FROM
       video_games
GROUP BY Platform
ORDER BY TotalSales DESC

/* RESULT ---------------------------------------------------------------------
Platform                    TotalSales
PS2                         1255.8
X360                        971.4
PS3                         939.6
Wii                         907.5
DS                          806.4
*/
```

```
/* Business Question 2:

   What is the trend of games released for given genre? Which platform had maximum
   games released on it?
--------------------------------------------------------------------------------------
   It could be measured as following:
*/

-- The Number of games released for given genre to know the trend
SELECT
       Genre
       ,COUNT(*) AS Number_Of_Games_Released
FROM
       video_games
WHERE Genre IS NOT NULL
GROUP BY Genre
ORDER BY 2 DESC

/* RESULT ----------------------------------------------------------------------------
Genre                Number_Of_Games_Released
Action               3370
Sports               2348
Misc                 1750
Role-Playing         1500
Shooter              1323
Adventure            1303
Racing               1249
Platform             888
Simulation           874
Fighting             849
Strategy             683
Puzzle               580
*/

-- Which platform had maximum games released on it?
WITH
cte_PlatformRank AS
(
       SELECT
              Platform
              ,COUNT(*) AS Number_Of_Games_Released
              ,DENSE_RANK() OVER (ORDER BY COUNT(*)DESC) AS Ranking
       FROM video_games
       GROUP BY Platform
)
SELECT
       Platform
       ,Number_Of_Games_Released
FROM cte_PlatformRank
WHERE Ranking=1

/* RESULT ----------------------------------------------------------------------------
Platform      Number_Of_Games_Released
PS2                  2161
*/
```

Q2. Games with multiple consoles:

a. How many games have been released with 3 or more Platforms?

b. In which year were the highest number of Genres at their peak ?
Please find <u>the Year & The Genres</u>

A2.

```
-- ANSWER 2a.------------------------------------------------------------------------
WITH
cte_GamesWithMoreThan3Platforms AS
(
       SELECT
        Name
       ,COUNT(Platform) AS NumberOfPlatformsByGame
       FROM video_games
       GROUP BY Name
       HAVING COUNT(Platform) >= 3
)
SELECT
       COUNT(Name) AS NumberOfGamesWithMoreThan3Platforms
FROM cte_GamesWithMoreThan3Platforms

/* RESULT --------------------------------------------------------------------
   NumberOfGamesWithMoreThan3Platforms
   1283
*/

-- ANSWER 2b.------------------------------------------------------------------------
WITH cte_GenrePeakYears AS (
    SELECT
        Year_of_Release
        ,Genre
        ,COUNT(*) AS Releases
        ,RANK() OVER (PARTITION BY Genre ORDER BY COUNT(*) DESC) AS Rank
    FROM
        video_games
    GROUP BY
        Year_of_Release, Genre
)
SELECT TOP 1
    Year_of_Release,
    COUNT(Genre) AS Peak_Genres_Count
FROM
    cte_GenrePeakYears
WHERE
    Rank = 1
GROUP BY
    Year_of_Release
ORDER BY
    Peak_Genres_Count DESC

/* RESULT --------------------------------------------------------------------
       Year_of_Release                Peak_Genres_Count
       2008                           5
*/
```

Q3. Finding the middle within the dataset:

Weighted Average: Like an ordinary arithmetic mean (the most common type of average), except that instead of each of the data points contributing equally to the final average, some data points contribute more than others.

$$\bar{x} = \frac{\sum_{i=1}^{n} w_i \cdot x_i}{\sum_{i=1}^{n} w_i}$$

Average: summing the values divided by the members count.
Mode: the most common value within the dataset

Calculate the weighted average, normal Average, and the mode of *critic_score* per rating. Please present all numbers rounded with 1 decimal point.
Which two *rating*s have the same values for all three measures? Please explain why

---

A3.

```
WITH cte_ScoreData AS (
    SELECT
        Rating
       ,Critic_Score
       ,Critic_Count
    FROM
        video_games
    WHERE
        Critic_Score IS NOT NULL AND Rating IS NOT NULL
),
cte_WeightedAverage AS (
    SELECT
        Rating
       ,ROUND(SUM(Critic_Score * Critic_Count) / SUM(Critic_Count), 1) AS Weighted_Avg
    FROM
        cte_ScoreData
    GROUP BY
        Rating
),
cte_NormalAverage AS (
    SELECT
        Rating
       ,ROUND(AVG(Critic_Score), 1) AS Normal_Avg
    FROM
        cte_ScoreData
    GROUP BY
        Rating
),
cte_ModeScores AS (
    SELECT
        Rating
       ,Critic_Score
       ,COUNT(*) as ScoreCount
       ,RANK() OVER (PARTITION BY Rating ORDER BY COUNT(*) DESC) as Rank
    FROM
        cte_ScoreData
    GROUP BY
        Rating, Critic_Score
),
cte_ModeResult AS (
    SELECT
        Rating
       ,Critic_Score AS Mode_Score
    FROM
        cte_ModeScores
    WHERE
        Rank = 1
)
SELECT
    a.Rating
    ,a.Weighted_Avg
    ,b.Normal_Avg
    ,c.Mode_Score
FROM
    cte_WeightedAverage a
INNER JOIN
    cte_NormalAverage b ON a.Rating = b.Rating
INNER JOIN
    cte_ModeResult c ON a.Rating = c.Rating

/* RESULT --------------------------------------------------------------------

Rating Weighted_Avg  Normal_Avg    Mode_Score
AO     93            93            93
E      73.3          68.5          70
E10+   71.4          66.8          73
K-A    92            92            92
M      75.2          71.8          84
RP     62.2          62            58
RP     62.2          62            63
RP     62.2          62            65
T      72.3          68.8          71

Two ratings that have the same values for all three measures are:
1) AO (Adults only) - 93
2) K-A (Kids to Adults) - 92
The reason all 3 measures of these ratings are the same is because they appear in the
dataset only once:
AO rating is seldom because of his restricted commercial availability -
publishers would edit the game to meet the M rating instead of keeping the AO rating.
K-A rating was changed in 1998 to E (Everyone).
*/
```

Q4. Data Scaffolding:

Please provide the global sales by genre, Platform, and Year.

Remember: Some of the combinations in between do not exist (such as for Platform '2600' for Action genre, the years 1984-1986 lack in the data – use the query below to validate that).

```
SELECT DISTINCT Genre, Platform, Year_of_release
FROM video_games
ORDER BY 1, 2, 3
```

You are required to display the measure for all possible combinations that can be between the fields (excluding NULLs) and bestowing zero when it's NULL for the measure.

A4.

```
/*
Query steps:
1) DistinctValues CTE: generating all possible combinations of genre, platform, and
year from the dataset, excluding NULLs.
2) SalesData CTE: aggregating the total global sales for each existing combination of
genre, platform, and year.
3) Final SELECT: joining these combinations back to the sales data. Where sales data
is missing for a combination, it defaults to zero using ISNULL.
*/

WITH cte_DistinctValues AS (
    SELECT DISTINCT
        g.Genre
       ,p.Platform
       ,y.Year_of_Release
    FROM
        (SELECT DISTINCT Genre FROM video_games WHERE Genre IS NOT NULL) g,
        (SELECT DISTINCT Platform FROM video_games WHERE Platform IS NOT NULL) p,
        (SELECT DISTINCT Year_of_Release FROM video_games WHERE Year_of_Release
         IS NOT NULL) y
),
cte_SalesData AS (
    SELECT
        Genre
       ,Platform
       ,Year_of_Release
       ,SUM(Global_Sales) AS Global_Sales
    FROM
        video_games
    WHERE
        Genre IS NOT NULL AND Platform IS NOT NULL AND Year_of_Release IS NOT NULL
    GROUP BY
        Genre, Platform, Year_of_Release
)
```

```
SELECT
    dv.Genre
    ,dv.Platform
    ,dv.Year_of_Release
    ,ISNULL(sd.Global_Sales, 0) AS Global_Sales
FROM
    cte_DistinctValues dv
LEFT JOIN
    cte_SalesData sd ON dv.Genre = sd.Genre AND dv.Platform = sd.Platform AND
    dv.Year_of_Release = sd.Year_of_Release
ORDER BY
    dv.Genre, dv.Platform, dv.Year_of_Release

/* RESULT ---------------------------------------------------------------------

Genre  Platform      Year_of_Release             Global_Sales
Action 2600          1980                        0.34
Action 2600          1981                        14.79
Action 2600          1982                        6.5
Action 2600          1983                        2.86
Action 2600          1984                        0
Action 2600          1985                        0
Action 2600          1986                        0
Action 2600          1987                        1.11
Action 2600          1988                        0.23
Action 2600          1989                        0.48

14.508 rows

*/
```

Q5. Year over Year analysis (aka: YoY)

Analyse per platform the year with the highest YoY % (Year of Year relative growth equation > (a – b) / b), in terms of *Global_Sales*.

Which of the following had recorded the most significant growth rate within the dataset, and in which year?

<u>Note</u> –

- In your analysis, take in account from 2nd year *Global_Sales* per *Platform* since the 1st year does not genuinely have a YoY % value.
- Same Data Scaffolding technique should be used here.
- Exclude 2020 from this data set.

A5.

```
WITH cte_AnnualSales AS (
    SELECT
        Platform
       ,Year_of_Release
       ,SUM(Global_Sales) AS Total_Global_Sales
    FROM
        video_games
    WHERE
        Year_of_Release IS NOT NULL AND Year_of_Release != 2020
    GROUP BY
        Platform, Year_of_Release
),
cte_YoY_Growth AS (
    SELECT
        a.Platform
       ,a.Year_of_Release
       ,a.Total_Global_Sales
       ,LAG(a.Total_Global_Sales) OVER (PARTITION BY a.Platform
        ORDER BY a.Year_of_Release) AS Previous_Year_Sales
       ,(a.Total_Global_Sales - LAG(a.Total_Global_Sales) OVER
        (PARTITION BY a.Platform ORDER BY a.Year_of_Release)) /
        LAG(a.Total_Global_Sales) OVER (PARTITION BY a.Platform ORDER BY
        a.Year_of_Release) * 100 AS YoY_Percentage_Growth
    FROM
        cte_AnnualSales a
)
SELECT TOP 1
    Platform
   ,Year_of_Release
   ,YoY_Percentage_Growth
FROM
    cte_YoY_Growth
WHERE
    YoY_Percentage_Growth IS NOT NULL
ORDER BY
    YoY_Percentage_Growth DESC

/* RESULT --------------------------------------------------------------------

Platform      Year_of_Release             YoY_Percentage_Growth
GBA                  2001                        87800
*/
```