1 Brainstorm

1. Box plot for continues columns (Sales, critic score, user score) ✅
2. Sales score by age rating.
3. Critic score by platform. ✅
4. Top 10 publishers by global sales ✅
5. Top 10 genres by global sales. ✅
6. Pie chart of the top 5 most sold genres and everything else is grouped in others. ✅
7. Pie chart of total Games sold by region ✅ (Hans)
8. Create a heat-map to compare the distribution of sales of the most popular genres based on different regions. (Michael)
9. Which genres sells the best on different platforms ✅
10. Compare critic_score/user_score of games with the same title but different platforms.
11. Games released over time grouped by genre. Showing the average user_score and critic_score on each axis, with a size being the number of games released. ✅
12. Which genres does each publisher prefer.
13. show evolution in game releases by age rating over time in a line graph. X axis is time, y axis is amount of games, with a line for each age rating category. see slide 24 from lec 11 for example

user satisfaction over time, trends in development.

1.1 Noter

- 6 Columns of data have a high procentage of null values. And will be handled case by case
- Nogle spil er dublikeret over flere platforme. Så nogle spil skal ligges sammen.
- Det er mulighvis ikke en god idé at kigge på spil solgt over tid, da vi kun har det totale antal spil solgt og udgivelsen af spillet.

Link to slides:
https://docs.google.com/presentation/d/1FS7jyo-yJZQmzOavr4HI-KJ4j3JnlbGyHPEsSUfRx94/edit?usp=sharing

1.2 How should we do Animations?

https://vis.csail.mit.edu/pubs/animated-vega-lite/

1.3 How should we do Interactive graphs?

Vega lite has it build in, see https://vega.github.io/vega-lite/examples/#interactive

1.4 Distribution of work between us.

Project Rapport

Hans Askov
haask19@student.sdu.dk

Michael Haugaard Pedersen
micpe18@student.sdu.dk

Sandra Malling-Larsen
small11@student.sdu.dk

2 Abstract

Briefly describe the data and what were the main achievements of your visualizations

Name	Percentage of Contribution
Hans Askov	
Michael Haugaard Pedersen	
Sandra Malling-Larsen	

Contents

1 Brainstorm 1
1.1 Noter 1
1.2 How should we do Animations? 1
1.3 How should we do Interactive graphs? 1
1.4 Distribution of work between us. 1
2 Abstract 2
3 Background and Motivation 4
4 Project Objectives 4
5 Data 4
5.1 Where to find Dataset 4
5.2 Description 4
5.3 Data Processing 6
6 Visualization and Dashboard 6
6.1 Design 6
6.1.1 Design of Graph 9 7
6.2 Must-Haves 7
6.3 Optional features 7
7 Story and Results 7
8 Conclusion and Discussion 7
9 Appendix 7

3 Background and Motivation

Discuss your motivations and reasons for choosing this project, especially any background or research interests that may have influenced your decision.

Video games have become a ubiquitous means of entertainment, both for children and adults, and have the unique ability of offering the user multi-dimensional enrichment by combining the storytelling of books and movies with the interactivity of traditional games and puzzles. Since the release of what is considered to be the first video game in 1958(1), there has been great growth in technology in general, and with it things like televisions and computers have gone from being a luxury to everyday object expected to be present in every home. Along the same line, video games have gone from being limited to big arcade machines to being able to run on most regular computers or smaller home consoles easily kept in a private home. Due to the evolution in technology, video games have been able to grow from simple pixelated graphics with limited button inputs to high definition photorealistic graphics with multiple forms of interactivity, ranging from button and mouse input, to touch screens, to motion sensor and alternate reality. With this, the audience for video games have also evolved, from being mostly aimed at children and family fun nights, to spanning many genres and all age groups. Knowing this, it would be interesting to explore the evolution of video games, to see if there are any clear trends in the kind of games developed, the target audiences for video games, and whether one gaming platform is able to rise above the rest and dominate the video game market.

4 Project Objectives

Provide a list of questions (and sub questions if relevant) which you plan to answer with your visualizations. What would you like to learn and accomplish?

The intention with this analysis is to explore the evolution in video games and to see if it is possible to identify any trends over time, it is not to answer a specific question like what is the perfect video game. The main areas to investigate are listed below.

- Sales
  - most sold genres
  - most sold publisher
  - distribution of sales by geographical region
  - distribution of sales by age rating
- Reviews
  - user scores vs critic scores
-
  - platform preferences

5 Data

5.1 Where to find Dataset

The dataset, Video Game Sales with Ratings up to 22/12/20216 - can be found on Kaggle.com via the link: Video Game Sales with Ratings

5.2 Description

describe all the relevant variables, number of records and any special feature of your data (if there is any)

The data set contains 11.563 unique entries, 16.719 total entries, and 16 variables, a summary of which can be seen in Table 1.

Name of Column	Data Type In Data Set	General Data Type	Null Value %
Name	VARCHAR	Categorical Nominal	0%
Platform	VARCHAR	Categorical Nominal	0%
Year-Of-Release	VARCHAR	Numerical Discrete	0%
Genre	VARCHAR	Categorical Nominal	0%
Publisher	VARCHAR	Categorical Nominal	0%
NA_Sales	Double	Numerical Continuous	0%
EU_Sales	Double	Numerical Continuous	0%
JP_Sales	Double	Numerical Continuous	0%
Other_Sales	Double	Numerical Continuous	0%
Global_Sales	Double	Numerical Continuous	0%
Critic_Score	BIGINT	Numerical Continuous	51%
Critic_Count	BIGINT	Numerical Continuous	51%
User_Score	BIGINT	Numerical Continuous	54%
User_Count	BIGINT	Numerical Continuous	40%
Developer	VARCHAR	Categorical Nominal	38%
Rating	VARCHAR	Categorical Ordinal	40%

The Name variable is the title of the video game, Year-Of-Release is when the video game was released, and Platform is the gaming system the game was released to, this includes both traditional console gaming systems, such as the XBOX and PlayStation families, as well as handheld consoles like the Nintendo Game Boy family.
For the last 20+ years, it has become pretty common for the same game title to be released for multiple platforms, eg. PC, XBOX, and Playstation) at the same time, which is why the same title might appear multiple times in the data set, with one entry per platform, leading to the difference between total entries and unique entries.

Genre categorizes the video games into one of 12 types of games; Action, Adventure, Fighting, Misc, Platform, Puzzle, Racing, Role-Playing, Shooter, Simulation, Sports, and Strategy.

Publisher is the name of the company that published the video game, which may or may not be the same company that developed the game. Likewise, the name of the developing company can be found in the Developer column. 38% of the games in the data set do not have a developer listed. This could be because the information is no longer available, due to companies no longer existing, or that the game was an indie development.

The Rating column contains the age rating the video game has received using the ESRB rating system, with the categories Rating Pending (RP), Kids to Adults(K-A), Everyone (E), Everyone 10+ (E10+), Teen (T), Mature 17+ (M), and Adults Only 18+ (AO), here listed from lowest to highest. In 1998 the rating K-A changed name to E.

The Critic_Score and Critic_Count columns show the aggregate score given by game critics and the number of scores used, respectively. These results were compiled by the staff of Metacritic, a website specializing in collating critic and user scores for various entertainment media. As this dataset spans

game releases going back to 1980, there are some games that Metacritic has not been able to combine critic scores for, simply because these critic reviews are not available online.

Similarly, the User_Score and User_Count columns show the average score given to a specific game by Metacritic's user base, and the amount of users that gave a score. As these results are only from Metacritic's subscribers, there is bias in the data set, both due to its incompleteness, but also due to Metacritic's user base not necessarily being representative of the users of the game titles.
Like with the critic scores, there are games that have no user scores, either because the game titles are old, or because the people who played the game aren't users of Metacritic; this can be due to lack of awareness, lack of interest in using the platform, or inability to use due to language barrier.
Additionally, Metacritic does not require any form of check or verification that a user has actually bought or played the game before giving a review score, meaning it is possible to artificially inflate or deflate the user review score.

The NA_Sales, EU_Sales, JP_sales, Other_Sales, and Global_Sales contain the total number of units sold in each region, in millions of units, from the release of a game up to 22/12/2016.
NA_Sales covers North America, EU_Sales covers the European Union, JP_Sales covers Japan, and Other_Sales contains the sales numbers from the rest of the world not covered by the previous three regions, ie. Africa, Asia minus Japan, Australia, Europe excluding the EU, and South America.
It is unclear if, and if yes, how, the dataset accounts for changes in European Union membership status, or if the sales numbers are based on the EU membership list of 22/12/2016.
Lastly, the Global_Sales column contains the sales numbers for a given video game's world wide sales.

5.3 Data Processing

Did you need to do substantial data cleanup? If yes, what techniques did you use?
Some of the variables have a high null percentage, meaning we are missing certain information for some of the data points. As can be seen in Table 1, this mainly affects the variables User_Count, Critic_Score, Critic_Count, Rating, User_Score, and Developer.
Seeing as rate of null values for the affected variables is so high, and there are questions where this information is relevant, the decision has been made to remove any incomplete data points from the data set. This does massively reduce the number of entries in the data set, from 16.719 total and 11.563 unique entries to 6.825 total and 4.449 unique entries, however, this exclusion allows for cleaner visualizations compared to the alternative.

Another change made to the data set is changing the K-A rating to E, for all affected columns. As described previously, the two ratings are effectively the same, with E replacing K-A. This change in the data set serves two functions; firstly it will allow the visualizations to more accurately depict reality by having one representation for the same age rating instead of two, and secondly, using only E will eliminate possible confusion for the reader, as most are no longer familiar with the K-A rating name.

6 Visualization and Dashboard

6.1 Design

How will you display your data? Provide some general ideas that you have for the visualization design. Describe your designs and justify your choices of visual encodings. sales distribution is heat map, not geo map, as the regions are not logically seperated(other contains areas that are not geographically close)

Sales over time grouped by genre is AI generated!!! we need to suggest alterations/make a better version

6.1.1 Design of Graph 9

We decided to use a heat-map to visualize which genres are popular across platforms. The data is normalized in regards to each platform, which is done to to allow for even comparison. The genres are oriented from the most popular genre to the least popular genre across all platforms, which means Action is at the top and most, where strategy is the least popular in the bottom. Intuitively this makes sense as being on top is generally associated with "winning" or being the best. another benefit of ordering genres this way is that it creates a visual gradient, but if there are parts of the gradient which does not match, then it indicates these platforms have something different to them. For example, shooter games are not popular with Nintendo platforms, but platform games are very popular with Nintendo platforms.

6.2 Must-Haves

List the features without which you would consider your project to be a failure.

1. You must have at least three types of graphs (i.e barchart, timeseries plot or boxplots) and
2. at least one animated graph (using for example gganimate).
3. An AI-generated graph must be included in the dashboard with description on how (from what platform and with what prompts) it is produced. Also you must specify which question regarding the dataset it answers.
4. In total at least 9 graphs. Provide clear and well-referenced images showing the key design and interaction elements.
5. A link to the dashboard/Visualization must also be included in the report.
6. An option to download the report as a manual from the dashboard

6.3 Optional features

List the features which you consider to be nice to have, but not critical.

7 Story and Results

here you should provide answers to the questions in 2.Project Objectives. Tell the story of the data that you saw in the visualization. What were your expectations and how close they were to what data revealed. What did you learn about the data by using your visualizations? How did you answer your questions? How well does your visualization work, and how could you further improve it?

By excluding incomplete datapoints, we are also excluding many of the games released in the first decade of video game releases. This means that many of the initial gaming platforms will not be a part of this exploration, which is a bit of a double-edged sword. On the one hand, this means there is less noise in the data when looking at recent developments in trends. However, at the same time, this lack of data from the infancy of home gaming consoles makes it very difficult to obtain any insights in the evolution of video games, and limits the possibilities for deeper investigations. Obsolete home consoles like Sega or Atari might not be relevant when looking at current market shares and trends, but would be vital if investigating why certain platforms flourished and other faded into obscurity. Granted, such an exploration would require additional data, such as console specifications and limitations, and could not be performed exclusively using this dashboard, but the lack of data from that time period prevents this dashboard from being useful for such an investigation.

8 Conclusion and Discussion

finally you conclude the report by a summary of what you achieved, how you achieved, what were the challenges for you and how the course can be improved.

9 Appendix