Nick Mavromatis (nima6629)

Storytelling Visualization Project Writeup

Video Game Sales Data Analysis

Introduction:

The project domain is that of video game sales of greater than 100,000 copies. This subject is important to inform future production of games, which is a very expensive business pursuit, so one would want to maximize the chances of a game's success. The average triple A game costs about 80 million dollars to produce. Additionally, the video game market is huge, being estimated at almost 200 billion in 2021 by sales, so it is very relevant to the business world. This project and visualization framework would be useful to a video game executive making decisions about future development, but it could also appeal to the lay person who is curious about the business of games, or even to an investor looking to capitalize on the market.

Visualizations are extremely powerful and useful in analyzing and exploring data trends. As demonstrated by the first assignment in this class where we attempted to summarize trends just by looking at an excel sheet, data can be very abstract and complex. It was nearly impossible to draw any useful insights or make interesting comparisons or summaries just by looking at abstract data. The power of visualizations is they allow the user to identify new patterns and make both broad and narrow assessments about the data and trends by looking at the data in an easier to grasp pictorial form. This allows the user to answer specific questions by interacting with the data. The interaction aspect is also key-a user can filter and slice data in a way to cut out the clutter and look at an attribute or trend of interest. Furthermore, different visualizations are easily linkable so that the connections between different aspects of the data become apparent. The user is thus able to perform many key tasks in an easy, streamlined way.

As described in Visualization Analysis and Design by Tamara Muzner, the guiding book of this course and project, vision is an extremely important sense to appeal to. Vision is a high bandwith channel of information that is particularly well suited to the display of information. The human brain is able to process a large amount of visual information in parallel at the preconscious level.

Sound, in contrast, is experienced as a sequential stream, and is poorly suited for analyzing large amounts of information in tandem. However, the video presentation uses an effective combination of both vision and sound to tell a cohesive story. Visualizations are also effective in showing data in detail, rather than just providing statistical figures without context. Interactivity is useful in mitigating complexity, and allowing investigation at multiple levels of detail. It can also mitigate the issue of limited screen real estate by dynamically changing views. Visualization idioms also focus heavily on tasks, which is useful in informing what actions the user will be able to perform and tailor the experience toward those requirements. For all these reasons, visualizations are an extremely useful scientific tool that can inform business analysis.

There are many interesting questions that can be answered and useful insight that can be gained by visualizing this dataset. After using this visualization system, the user could understand which video games sold the highest (or lowest) by analyzing a variety of attributes such as global sales by genre. An investor could also look at global sales by publisher to assess which publishers to potentially invest in for the greatest return. A visualization of global sales by platform could also provide insight into which platforms are expected to have the greatest sales. A graph of global sales by publisher and genre could provide insight for any type of user into what genres a particular publisher specializes in or excels at, to guide production or investment in the future. Visualizations of sales for different regions by genre and publisher would provide insight into which regions to target specific genre releases to. Finally, the trend of global sales by year could also be powerful in making future predictions.

I chose to implement a storytelling project, focusing on creating a cohesive story from the data. There are already so many great visualization software packages, that I think it makes more sense to focus on learning them rather than reinventing the wheel by trying to implement a programming implementation with greater limitations. I normally enjoy the experience I get from programming assignments, so this was a difficult choice, but I feel that gaining more skills using existing visualization software will be the most helpful in real life contexts. I know that Power BI is heavily used in the industry, and has a significant learning curve, so perfecting its use is a worthwhile task. I was able to draw many interesting conclusions using Power BI visualization.

Dataset:

The data set was collected from Kaggle.com found in Google Dataset Search, and was compiled by Gregory Smith by scraping vgchartz.com. The data is in the format of a spread sheet with over 16,500 entries from the years of 1980-2016. Name is a straightforward categorical attribute, containing the title of the game sold. Genre and publisher are also categorical attributes. Rank is organized according to global sales and is an ordered quantitative sequential attribute. NA sales, EU Sales, JP Sales, other sales, and global sales are all ordered quantitative sequential data types and the scale is in millions of copies sold. The sales are hierarchical, in that the other four categories make up global sales. Year is a temporal attribute, corresponding to the release of each game. The keys of table are either rank or name, which are unique to each entry.

Overall the data quality is very high. There are very few missing or incomplete entries in the dataset. There are about 200 entries that do not have the year, so I filtered these out when doing analyses based on year. However, there are over 16,000 entries in total, so that isn't a significant loss. There are also less than ten entries each for the years 1980, 2017, and 2020, so I filtered these out when doing temporal analyses, so it didn't present a skewed view of the profit or genres corresponding to those years. It is very easy in Power BI to filter out certain values, so rather than delete the data entirely I elected to filter the data for temporal analyses. The data is still fine when considering analyses that do not include year.

Tasks:

There are multiple domain related tasks that a user of the system will be able to perform to better understand trends in the data. A user will be able to see the total sales organized by genre in order to predict which genres will sell well or poorly in the future. The user will also be able to analyze sales by publisher to predict which company to potentially invest in to maximize returns. The user can also see sales by region, of the most successful games organized by genre, to better understand which markets to focus on and where to target specific genres. A user is also able to analyze sale trends by year and platform to summarize trends in the industry. The specific tasks that support these goals are to analyze and consume the data to answer specific questions, and to search the data in order to discover trends. To these ends, the dashboard allows slicing of the data, by attributes such as number sold, region, genre, or publisher and it allows filtering by specific values of interest. This is seamlessly done by clicking on a particular attribute of a visualization in power BI, which synchronizes the linked view across multiple visualization idioms automatically. For example, the user can click on a publisher in one visualization to filter the other graphs by that publisher, showing the most successful sales of all times, trends over time, and sales by region among other information.

There are also more abstract task actions and goals that guided the visualization system. The system is mainly geared toward experts looking to guide future development or investment decisions, but a lay person can also use the tools easily. The main goal is for the user to be able to analyze larger trends in the industry by consuming the information presented in the visualizations. Discovery of new insight into the video games market is another important task abstraction, which can be used to form and validate hypotheses. I had some hypotheses before undertaking this project, and some of them were confirmed. Also, a user can generate a new hypothesis, which is a powerful application. Finally, the visualization system of power BI allows the user to present an interesting and cohesive story that could capture the audience's attention. This is done in the presentation by filtering on specific attributes and values in order to change linked views and drill down to greater detail. To support these tasks, the user can search for data by lookup, locating, browsing, and exploration, as well as performing more specific queries. With these actions, I was able to compare different attributes, and summarize larger trends of the industry, which may be invaluable to making future business decisions.

Visual Designs:

Power BI allowed for the creation of an integrated, interactive system of visualizations from which powerful insight can be drawn. I created a dashboard of several different visualizations of multiple attributes with different idioms. The key thing is that the user can click on an aspect of one visualization, such as genre, and then filter by this particular value across each of the visualizations in one integrated linked view. Through this linked navigation and selection, the result is faceting of the graphs by a specific element of interest and linked selection and highlighting across the many views. Importantly, the different views are juxtaposed side by side, but the user is also able to perform semantic zooming to focus on just one graph when the screen real estate would otherwise not be sufficient. This allows a high degree of control and freedom at both low and high levels of detail, which allows for an effective presentation of results. I chose to filter based on clicking on a graph rather than creating separate scented widgets in order to simplify ease of use and maximize the use of the screen space. Below, I explore each of the visualizations created with justifications for their inclusion.

The first visualization is a simple chart of the highest selling games in one column and total global sales in the other. This chart is so useful in that one can click on a particular genre, region, or publisher in a different graph to display the highest selling results in the chart. Sometimes, simply representing numbers in a non-abstract, literal way is the most relevant approach, whereas masking this visualization with other marks and channels would lead to less clarity and precision.

The second visualization is a simple column bar chart of global sales by genre. I chose a simple representation so as not to obscure the analysis of trends, and it also allows for the user to click on a particular genre and link and filter all the other views by this genre. I chose to go with the simple built in color scheme, leaving the bars as blue which is visually please and not distracting.

The next visualization is a pie chart of global sales by publisher. This allows interactive linked filtering by one particular publisher, which allows the user to easily compare the performance of different publishers to make potential investments. A pie chart was the idiom of choice because it allows the easy visual comparison of the number of sales between different publishers in a way that pops out effectively. Pie charts are effective in displaying a part to whole relationship in a qualitative way, but also display exact quantitative values of market share. Here, the user is focused on the highest performing publishers, in that the lowest performing publishers are hard to make out, but this is easy to change by semantic zoom. Most of the graphs are sorted by highest to lowest, because the highest selling attributes are of greater interest.

Another visualization is a horizontal clustered bar chart of sales by region by genre. I found that horizontal bars used the horizontal screen space more effectively, and clustered bars were more easily compared than stacked bars like I had originally proposed. This graph crucially allows the user to link and filter views by region, which can lead to new insight.

Another visualization is a line graph of global sales by year, organized sequentially by year. This is an effective representation when displaying temporal data and allows quick identification of the peak year of sales. The user can also filter the linked views by year if necessary.

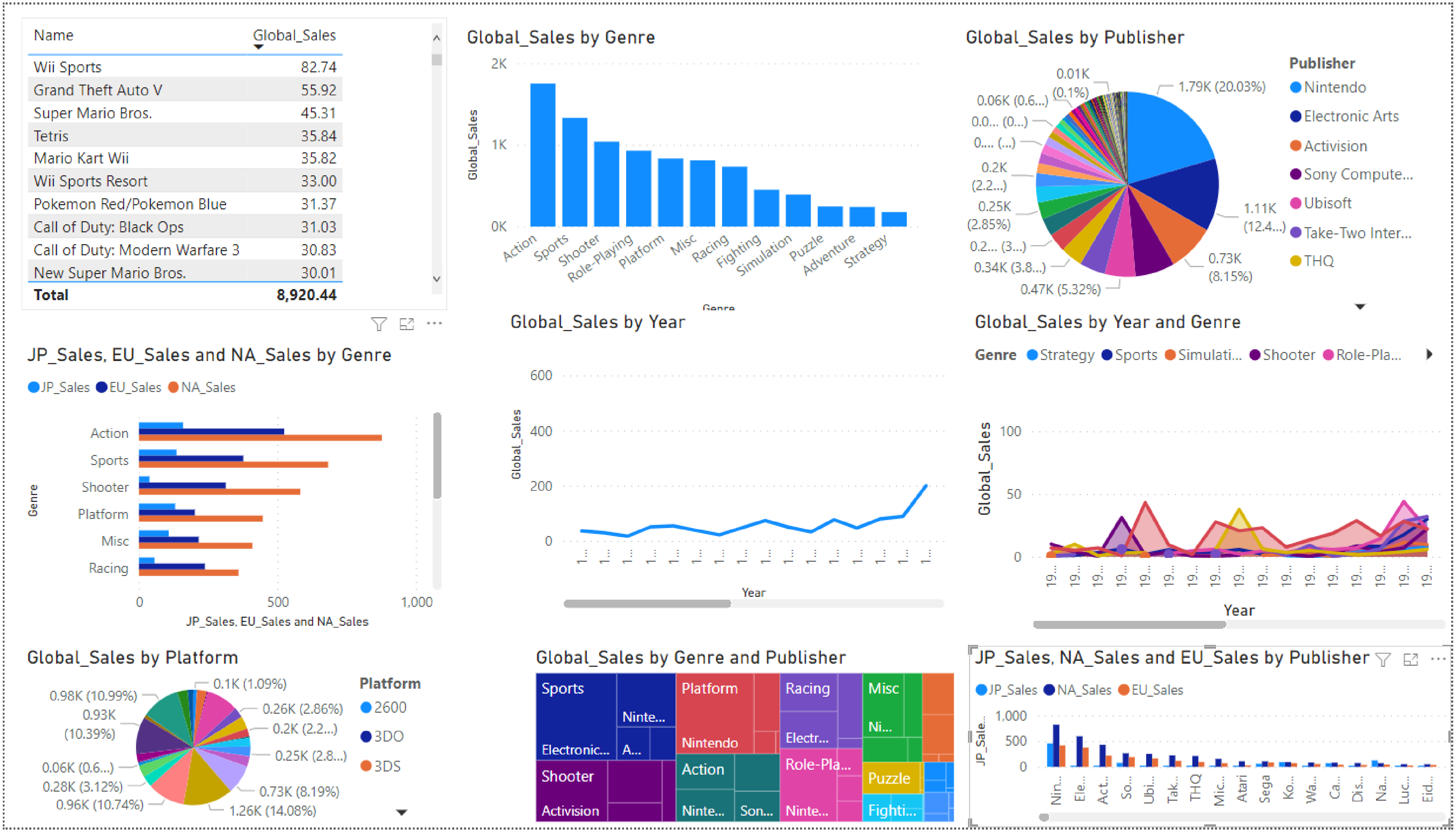
The next idiom chosen is an area chart of global sales by year and genre, which allows for analysis and comparison of trends in time over genre. This visualization is difficult to interpret until the user clicks on the tool bar key for a particular value, which allows selective highlighting by genre value and is a crucial interactive element. This is most effective after performing a semantic zoom, as the data is naturally spread out over a large horizontal axis.

The next visualization is a pie chart of sales by platform. This allows the user to draw inferences about which platforms were most appealing to the consumer, and could allow a hardware producer to make decisions about which platforms to produce in the future, which is an expensive but potentially lucrative venture. It also allows for the game production companies to choose which platforms to develop for, which can have a major impact on sales numbers. On a side note, I was struck by how much better primarily 3D systems sold than solely 2D systems. This was also sorted highest to lowest with the default color scheme, making analysis of part to whole relationships easy.

One of the more interesting representations is a hierarchical tree map of global sales by publisher and genre. This allows for cursory conclusions and comparisons to be drawn about which publishers dominate in which genre space, and clicking on one selection also populates the simple chart of highest selling games to allow the user to easily drill down into greater detail. A hierarchical representation was appropriate here because sales rank by genre is a naturally hierarchical category.

Finally, the last visualization is a vertically clustered bar chart of sales by region and publisher. This allows for the analysis of which publisher did best in which region. Semantic zooming is crucial for this representation, as there is too much information to be effective in the zoomed out state.

Overall, I mostly chose simpler idioms which are easier to interact with to represent the data. This allows for more effective conclusions to be drawn, as the learning curve for a new user is low. It also obscures the results less than a flashier choice would. However, the implementation of highly linked views allows for complex manipulation and greater insights to be drawn from a simpler set of base tools.



Name	Global_Sales
Wii Sports	82.74
Grand Theft Auto V	55.92
Super Mario Bros.	45.31
Tetris	35.84
Mario Kart Wii	35.82
Wii Sports Resort	33.00
Pokemon Red/Pokemon Blue	31.37
Call of Duty: Black Ops	31.03
Call of Duty: Modern Warfare 3	30.83
New Super Mario Bros.	30.01
Total	**8,920.44**

Figure 1: Power BI visualization dashboard, showing visualizations described in section above.

Results:

First, I will describe the process used to create the visuals and story, then move into key insights and results. I first considered the specific questions to be asked, such as “which genres sell the highest” and from these, constructed preliminary visualizations with the attributes set appropriately. For data involving time, line graphs were a natural choice. The view involving rankings of global sales by title naturally lent itself well to a simple chart without further abstraction or obscuration of the results. Data involving the comparison of multiple regions on one attribute were well fit by clustered bar graphs. I selected pie charts to display part to whole relationships, such as sales by publisher, which allows for quick comparisons. Global sales by genre and publisher was naturally hierarchical, so lent itself well to a tree graph. Finally, I decided that global sales by year and genre would be well represented by a stacked area chart which allows for easy filtering and drill down by clicking on the legend, and is effective when one wishes to track not only the total value, but also the breakdown of that total by groups. To create the story, I answered the questions I posed in the beginning and considered how interaction could visually demonstrate these findings.

There were many interesting insights gleamed from the visualization system. The highest selling games within the time frame of the study are Wii sports, Grand Theft Auto V, Super Mario Bros, and Mario Kart Wii in that order. There are also several Call of Duty games in the top ten list. It stands to reason that Mario and casual sports games are often very successful, but that among a more mature audience, shooters and simulation games are also popular.

Looking at global sales by genre, the best selling genres are action, sports, shooters, role playing, and platforming games. Puzzle, adventure and strategy games sold the worst, so may not be a good bet to invest in in the future. Nintendo, Electronic Arts, Activision, Sony Computer Entertainment, and Ubisoft are the best selling publishers, but Nintendo and Electronic Arts dominate with around 20% and 12.5 % of sales respectively. Clicking on each of the three greatest selling publishers, Nintendo specializes in sports, platforming and racing games. Electronic Arts specializes in sports games, while Activision is most successful selling shooters. The tree map also shows these findings in a nice way. These companies are usually solid bets to invest in, and are projected to continue to be successful into the future.

The global sales by year chart shows a peak of sales in 2008, likely boosted by the high sales of the Nintendo Wii, and blockbusters like Grand Theft Auto IV. However, game sales fell in the following years, likely due to the recession. Although the data isn't current, external data shows that the game and hardware sales market has been steadily rising and is expected to continue to rise in 2023.

Sales by region show that sales are far greatest in the U.S., which makes sense as many publishers focus on targeting this region. Sales in Europe are usually about 50% less, while sales in Japan, a low population country, are by far the lowest. Action, sports, and shooter games do particularly well in the U.S. and Europe, while role playing games are popular in Japan, which agrees with my previous knowledge and hypothesis. Puzzle, adventure, and strategy games do poorly in all three regions. The area chart of global sales by year and genre show that the highest selling and fasting growing genres are sports and shooters, likely mainly due to the U.S. and European markets.

Amongst the platforms, primarily 3D platforms have by far sold the best, including the Wii, Xbox 360, DS and PS3. Nintendo is the only company to sell high numbers of handheld systems. It seems that, despite increased complexity, 3D games appeal to purchasers far more than 2D games, perhaps due to more impressive visuals.

To summarize, sports, action, platforming, and shooters are the most popular genres. Nintendo, Electronic Arts, Activision and Sony dominate the market, each with a different focus on genres. Sales are far greatest in the U.S., with the largest gaming market, and about half as high in Europe and far lower in the less populous region of Japan. Sales peaked in 2008, fell for a few years due to the recession, but are now at an all time high. Shooters and Sports games are usually a solid bet and continue to rise in sales numbers. These findings could be invaluable to both game and hardware production companies, as well as those looking to invest in the potentially profitable market.