TAG-ORIENTED MARKET ANALYSIS OF ONLINE MARKETPLACES USING VISUAL ANALYTICS - “GAME MARKET VISUAL ANALYTICS”

Valerio Di Stefano (1898728) - Visual Analytics Final Project

***Abstract** - Market research, the market analysis methodology for business organizations, allows businesses to identify the needs, size and competition of a market. When market research is paired with information visualization and visual analytics methodologies, business insights become accessible to a wider range of users. The paper describes a tag-oriented visual analytics system for market research applied to online markets, which focuses on data about the online desktop game distribution platform “Steam”.*

1. INTRODUCTION

Market analysis is the activity of gathering information about conditions that affect a marketplace. It studies the attractiveness of a market, defined as the degree to which a market offers opportunities to an organization, taking into account the market size, its growth rate, the level of competition and other constraints.

For business organizations, market analysis is centered around market research, defined as the the action or activity of gathering information about consumers' needs and preferences, aimed at identifying the needs of the market, the market size and the competition. These kinds of business activities are usually conducted by data analysts and data experts within companies, relegating the understanding of its intricacies to a small portion of a specific organization, since the large amounts of collected and produced data might need a high level of expertise for its cleaning and transformation into actionable business decisions and insights. Furthermore, small organizations may not have enough human or capital resources to allocate for market analysis and market research, resulting in less informed business decisions due to the lack of dedicated personnel capable of correctly gathering and analyzing the large amounts of available data.

Employing visualization techniques and visual analytics solutions might enable individual entities, organizations and their employees, with a wide range of backgrounds, to view and understand market analysis data and its insights.

The business data at the center of market analysis and market research may be related to a very heterogeneous set of entities (data about sales, available resources, finance, customers, revenue or profit, and many more). The data-driven nature of modern business companies centered around these various classes of business data appears to be the perfect setting for applied visual analytics and information visualization approaches.

Despite this seemingly perfect fit for the visual analytics field, the literature on visual analytics solution for market analysis and market research staled for the last 6 to 7 years, while major transformations took place across different industries, influencing the way organizations conduct their business, most notably due to the Covid-19 pandemic of 2020, which significantly accelerated the digital transformation of business (among other aspects of the human life) [9]. This recent acceleration of the digital transformation led to a rapid shift in consumer behavior towards online shops, e-commerce and online markets, which quickly got used to buying digital but also physical products online. The transformation also led to a growth of e-commerce and online marketplace platforms (major tech companies like the american Amazon or the chinese Alibaba with their massive online marketplaces, or retail and physical stores proposing online alternatives to their physical stores), as well as digital goods and services marketplaces (such as mobile app stores like Android’s and Apple's, software and game digital stores, but also digital goods marketplaces and digital subscription services, like Netflix or Amazon Prime for movies). The panorama of visual analytics solutions for businesses did not grow with the same speed of the digital industries transformation, and the solutions and systems proposed before it only focus on traditional business practices centered around geographical distributions of shops and physical sales location [2], or on specific markets rather than broad consumer markets and large marketplaces for physical/digital goods and services [5].

The solution proposed in this paper aims at presenting a case study of the desktop gaming market, which was valued at 29.16 Billion U.S. dollars in 2022 and is expected to grow to 32.11 Billion U.S. dollars by 2030 [10]. Due to its large volume, the desktop gaming industry may represent a suitable case study to generalize to the consumer digital market as a whole, and to online markets in general, presenting various similarities in the way goods and services are exchanged and in the way market analysis for customers, competitors and products can be conducted.

The paper introduces a tag-oriented market analysis methodology aimed at identifying trends, growth possibilities, products supply and demand, and customer preferences and behaviors in the desktop gaming market.

The proposed solution, in particular, analyzes the online video game marketplace “Steam”, the biggest desktop-oriented game distribution platform with over 85.000 games and software products and over 130 million monthly active users as of the end of 2023 [11], with a 44% desktop gaming industry market share as of the end of 2022 [10].

The paper provides a description of the design process and rationale behind the proposed visual analytics solution for a tag-oriented market analysis. It highlights the employed datasets, analytics and implemented visualizations, the differences with respect to existing solutions in the literature, and also presents some use cases and insights resulting from the interaction with the system.

2. DATA

The work to implement the visual analytics solution described in this paper started from the gathering of data about the “Steam” marketplace. This data, as often happens with online marketplace data in general, was only comprehensive of publicly available information mainly directed towards consumers and customers, such as item description, features, price, characteristics, and general product information. Data that could be used for market analysis and market research, such as sales data or business data in general, were not directly available, and had to be estimated using existing techniques proper of the gaming industry ecosystem (described later in this section).

In particular, the starting dataset was collected from the “Kaggle” platform [6], and consisted of 84.368 unique data items with 39 attributes each, 10 of which were attributes with additional data structures as values (lists of single values, lists of key-value pairs or lists of complex items with more attributes).

An initial cleaning of the data was necessary (ad-hoc Python scripts were created for this purpose) to filter out any non-gaming item from the original dataset, which narrowed down the whole dataset to 73.078 game items.

Furthermore, various data attributes were simplified and cut out, and newly computed data was added (once again using Python). In particular, 12 non-relevant attributes were cut from the collection, and 4 of them were aggregated and simplified (from lists of complex objects to simple lists).

Two new attributes were computed for each item, and added to the collection: an estimate of the amount of copies sold of the game (computed using the “Boxleiter method” [12], a well known method in the desktop gaming industry for estimating game sales from the number of game reviews), and an estimate of the revenue of the game (based on the copies sold estimate and the price of the game).

The final games dataset used in the system is composed of 73.078 items with 17 attributes each (6 of which representing lists of values): name, steam ID, developers, publishers, header image, release date (year, month, day), price, copies sold, positive reviews, negative reviews, languages, player mode, content rating, revenue and tags.

As mentioned in the introduction, the solution proposed in this paper is centered around "tags", which can be described as publicly visible text labels that game publishers and customers of the Steam platforms can assign to each game. Up to 20 tags chosen from a pool of 448 tags can simultaneously be assigned to a single game item. At the moment of publishing a game to the steam platform, the publisher can choose between 20 of these available tags to assign to their published game, and after the game's release, users of the Steam platform can propose changes to this list by suggesting to add or remove certain tags. The tags assigned to each game item are also ranked by relevance (initially decided by the game publisher, then updated based on users' votes on the relevance of each assigned tag).

The original dataset included the list of tags associated to each game as key-value pairs representing the tag name and associated relevance among the game's tags list respectively. The final dataset simplified the list of tags of each game to a simple list of names, but the available tag's relevance values were used to compute a data set of tags of the Steam platform. The total number of 448 tags was narrowed down to 423 by excluding items not related to games, and a set of attributes was assigned to each tag item to produce a separated dataset for tags' information. These attributes include the name of the tag, the number of games which have that tag, the total global revenue of all games with that tag, as well as the total copies sold of games having that tag.

Other than these attributes, a new "category" attribute was also computed to classify the tags into three categories: "Genres", "Sub-Genres" and "Features": the classification derived from an initial (incomplete) classification of tags that the Steam platform implements, found in the documentation for Steam partners [13]. A final additional set of attributes were generated for each of the tag, representing the coordinates of both the t-SNE (t-Distributed Stochastic Neighbor Embedding) and MDS (Multi-Dimensional Scaling) dimensionality reduction techniques used for tags similarities (described later in the paper).

Starting from the newly derived tag items collection and from the cleaned and expanded game items collection, a third, additional dataset was built, to store the "similarity values" among tags used throughout the system's visualizations. This "similarity value" was computed for each pair of different tags as the number of times one of the two tags appears with the other tag in a single game item, then normalized by dividing it by the number of occurrences of the tag with the lowest number of games among the pair. The similarity values among pairs of tags were stored in a matrix-like data structure used in the system.

The final dataset used for the application was also composed of several lists of games and tags (represented by their IDs) indexed by features and attributes criterias, to be used for the various system's ranking visualizations, and thus to avoid having to sort filtered data items (both games and tags results) by said attributes or features.

All of the described datasets were built "offline", by computing their values using Python and storing the collections in JSON files to be used by the Javascript code of the web app that represents the system.

3. VISUALIZATIONS & ANALYTICS

The system's user interface, as presented in Figure 1, is composed of 10 coordinated views, which implement different types of visualizations of different types of data, with various types of possible interactions to help with the analysis phase.

Some of the views of the systems are highly customizable, and thus can change the correlated visualizations drastically, sometimes producing a different visualization entirely (it's the case, for example, of the tags scatterplot, which becomes a bubble chart when an attribute to plot is chosen for the points' radius encoding, or of the similarity visualization, which allows to switch between an MDS plot, a t-SNE plot, and a chord diagram). All the interactive visual elements of the system (points and bubbles of scatterplots, bars of histograms and rankings, cells of tables and treemaps, texts and labels, control buttons of each section, and more) also have a dedicated tooltip which appears on hover (that can also be disabled with the specific option) and provides an extensive description of the elements with which the user can interact, of their functions and of their visual encoding. Moreover, while the system's intended resolution is Full HD (or higher), each of its sections and views can be focused by clicking on the corresponding title, which will immediately scale it up to fill all the available screen space, and then can quickly be reset to its original size by clicking outside of the section itself or clicking on its title again (an option is also available to enable a scale up of each section on hover).

Finally, each view of the system also includes an info button (the "?" button on the top right corner of each view) which, on click, shows a "tooltip" to explain what the visualization represents, how its visually encoded data is computed and how to interact with its elements. The remaining part of this paper's section will briefly describe each visualization and the related analytical functions.

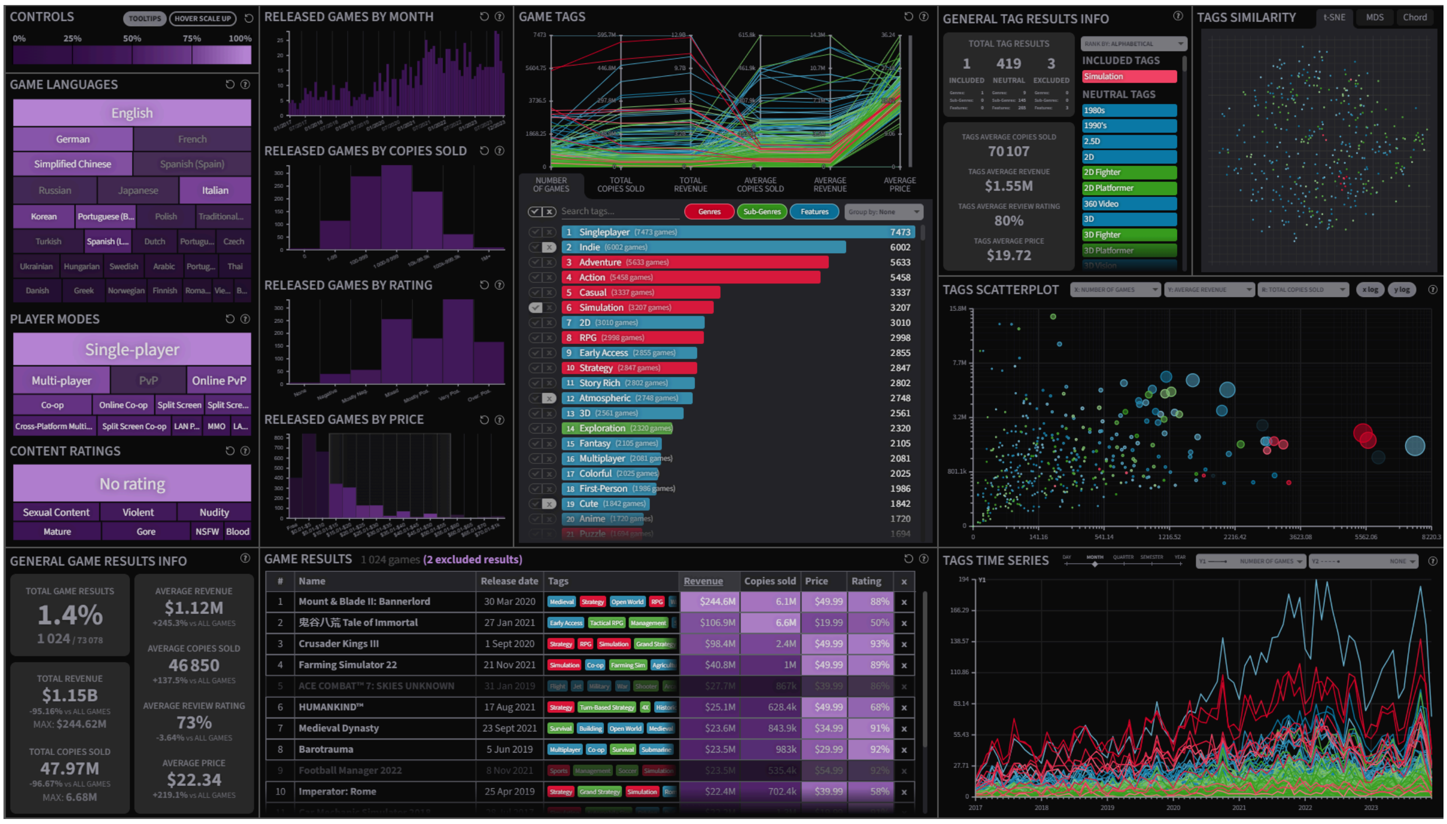
3.1. TREEMAPS

The first view of the system's visualization, from the top left of the user interface, is composed of three treemap visualizations which present information about game languages, game player modes and games' content ratings respectively. Each treemap cell represents one of the categories in these three attributes, and its size and color redundantly encode the number of games presenting that specific characteristic or feature. Redundancy was used because oftentimes treemaps can be hard to fit into a rectangular space without adopting some sort of distortion of the areas of each cell, and that's because of the need to present each cell in a horizontal orientation (to allow for the specific category name to be read) and the need to impose a minimum size to cells to ensure the visibility of even the ones which should actually not be visible if the encoding was strictly followed without any form of distortion or constraint. Employing a color scale for the background of each of the treemap's cells allows to distinguish between values of adjacent cells more easily than just using rectangular areas. The color of each cell is drawn from a gradient that can be seen on the top left of the UI (and also on hover of each cell) which follows a standard single-hue sequential color map obtained using the ColorBrewer system [8].

The reason for the need of keeping an horizontal orientation of the cells and for the need of ensuring that each cell is never smaller than a predefined minimum size is that each of the three treemap visualizations also acts as filters for the encoded game attribute, hence allow to exclude or include a certain game language, player mode or content rating from the results by clicking onto the cell to enable/disable it. On interacting with the cells of a certain treemap, the other visualizations of the system (including the other two treemap visualizations) will be updated to compute their data by considering only game results that have one or more of the attributes corresponding to each of the enabled treemap cells, as also happens with all the other visualizations' elements (all views are coordinated). Again, as for any other visualization's element, hovering over each treemap cell will provide additional information about the specific language, player mode and content rating.

3.2. HISTOGRAMS

The second view of the system, from the top left of the system's UI, is represented by four vertical histograms encoding: the number of released games over each month



#	Tag	Games
1	Singleplayer (7473 games)	7473
2	Indie (6002 games)	6002
3	Adventure (5633 games)	5633
4	Action (5458 games)	5458
5	Casual (3337 games)	3337
6	Simulation (3207 games)	3207
7	2D (3010 games)	3010
8	RPG (2998 games)	2998
9	Early Access (2855 games)	2855
10	Strategy (2847 games)	2847
11	Story Rich (2802 games)	2802
12	Atmospheric (2748 games)	2748
13	3D (2561 games)	2561
14	Exploration (2320 games)	2320
15	Fantasy (2105 games)	2105
16	Multiplayer (2081 games)	2081
17	Colorful (2025 games)	2025
18	First-Person (1986 games)	1986
19	Cute (1842 games)	1842
20	Anime (1720 games)	1720
21	Puzzle (1694 games)	1694

GAME RESULTS 1 024 games (2 excluded results)

#	Name	Release date	Tags	Revenue	Copies sold	Price	Rating	x
1	Mount & Blade II: Bannerlord	30 Mar 2020	Medieval Strategy Open World RPG	$244.6M	6.1M	$49.99	88%	x
2	鬼谷八荒 Tale of Immortal	27 Jan 2021	Early Access Tactical RPG Management	$106.9M	6.6M	$19.99	50%	x
3	Crusader Kings III	1 Sept 2020	Strategy RPG Simulation Grand Strategy	$98.4M	2.4M	$49.99	93%	x
4	Farming Simulator 22	21 Nov 2021	Simulation Co-op Farming Sim Agricultu	$40.8M	1M	$49.99	89%	x
5	ACE COMBAT™ 7: SKIES UNKNOWN	31 Jan 2019	Flight Jet Military War Shooter	$27.7M	867k	$39.99	86%	x
6	HUMANKIND™	17 Aug 2021	Strategy Turn-Based Strategy 4X Historic	$25.1M	628.4k	$49.99	68%	x
7	Medieval Dynasty	23 Sept 2021	Survival Building Open World Medieval	$23.6M	843.9k	$34.99	91%	x
8	Barotrauma	5 Jun 2019	Multiplayer Co-op Survival Submarine	$23.5M	983k	$29.99	92%	x
9	Football Manager 2022	8 Nov 2021	Sports Management Soccer Simulation	$23.5M	535.4k	$54.99	92%	x
10	Imperator: Rome	25 Apr 2019	Strategy Grand Strategy Simulation Rom	$22.4M	702.4k	$39.99	58%	x

Figure 1: System's user interface composed of 10 main visualizations (from left to right, top to bottom): Treemaps Sections (top left, for game languages, player modes, content ratings), Histograms Sections (top left, for released games by month, copies sold, rating, price), Parallel Coordinates View (top center), Tags Ranking (center), General Tags Infos (top right), Tags Similarity Visualizations (top right, including t-SNE, MDS, Chord Diagram), Tags Scatterplot / Bubble Chart (center right), General Games Infos (bottom left), Game Filter Results Table (bottom center), Tags Time Series (bottom right).

(from January 2017, year in which the Steam platform became open to external publishers and developers with the removal of the previous "Steam Greenlight" programme [14], to December 2023), the number of released games by total copies sold, then by review rating and lastly by price.

Each histogram presents its bins (or buckets) on the X axis, and its corresponding attribute on the Y axis, such that bars' height encodes the value for said attribute in the corresponding bucket. The color of each bar (which follows the single-hue gradient encoding used throughout all visualizations, as previously described) represents instead the percentage of filtered games that fall into that bucket with respect to all games that fall into the same bucket.

In a similar way to treemap cells, histogram bars act as filters that can be defined over the four attributes corresponding to each histogram: by clicking onto one of the histogram bars and dragging onto another bar, the active filters of the visualization will be set to only include games that fall into the specific range selected by the interaction.

Apart from the straight forward buckets definition of the "released games by month" histogram, the other histograms follow a specific distribution of buckets. The copies sold histogram follows a logarithmic scale for its buckets definition, which starts from the simple "0 copies" bucket to the "1-99" range, then the "100-999", then the "1000-9999", ecc…, up to the "1M+" bucket, which collects all games that sold more than one million copies (this logarithmic distribution reflects the underlying distribution of games' copies sold, which rapidly falls off as the number of copies sold grows). The review rating histogram groups bucket into the review ratings used on the Steam platform ("None", "Negative", "Mostly Negative", "Mixed", "Mostly Positive", "Very Positive", "Overwhelǹimingly Positive") each corresponding to a specific range of percentages for positive reviews [15]. Finally, the price histogram presents price ranges in chunks of five dollars differences, starting from the "Free" price and then continuing with "$0.01 - $5.00", "$5.01 - 10.00", ecc…, up until the last range of "$70.00 - $1000.00" dollar range ($1000.00 is the maximum game product's price allowed by the Steam platform).

3.3. PARALLEL COORDINATES VIEW

The parallel coordinates visualization of the system allows to visualize lines corresponding to the six main tag attributes computed by the system at each new filters update: number of games, total revenue, average revenue, total copies sold, average copies sold, average price.

At each update of the filters, the system goes through the list of games, identifying which of these games meet the currently active filter criterias, and for each game scans through the list of assigned tags and computes values corresponding to total and average values of the aforementioned tag attributes. These computed values are then used to plot the tag's parallel coordinates' lines (as well as other visualizations involving tags).

Tag lines are colored with one of three colors corresponding to the three possible tag categories ("Genres", "Sub-Genres" and "Features"), which are part of a categorical 4-colors palette (red, green, blue and purple) obtained from ColorBrewer's palette picker [8], consistent throughout all the visualizations of the system.

Colors of the tag lines (and in general of every other element corresponding to a tag in each visualization, like points, bubbles, lines and rectangles) are also offsetted in saturation and lightness to produce a single hue color scale that could allow tag elements with the same category (and thus the same color's hue), to be distinguishable in each visualization.

The tags' parallel coordinate view also shows a horizontal line on hover (with values for the corresponding coordinates of the six attribute axes) and allows for controlling the vertical scale of the view to implement zooming functionalities (paired with dragging functionalities to navigate the graph).

3.4. TAGS RANKING

Upon clicking onto one of the labels of the six axes of the tags parallel coordinate view, the user can choose to rank tags based on the corresponding attribute. The scrollable ranking shows, for each tag, a bar whose width is proportional to the ranked attribute's value, the tag name, the color corresponding to the tag category, the number of games with that tag that meet the current filter criterias, and the numerical value of the attribute chosen for the ranking. Hovering over each tag element also shows a tooltip with information about all of its attributes value, and highlights the corresponding tag line in the tags parallel coordinate view.

On the left of each tag, two buttons also allow to include and exclude a tag, or to reset its state to neutral: as also explained by the tags ranking visualization's info section, "including" or "excluding" a tag means filtering the game results visualized throughout the various system's visualizations to only take into account games with or without that specific tag in their associated tags list respectively. The tags ranking also allows to search for the name of a specific tag in the ranking, to quickly include and exclude all tags from the filters or all tags with a certain associated category, and to group tags by category or by their included/excluded state in the ranking, all by using the dedicated controls on top of the tags ranking view. Ultimately, brushing onto tags ranking bars allows to select a certain range of tags to quickly include/exclude (left click) or to include/exclude all tags above or below a certain selected tag in the ranking (right click).

3.5. GENERAL TAGS INFO

The general tags info section provides numerical values which correspond to the number of included, excluded or neutral filters for each tag category, and the global average values of copies sold, revenue, price and review rating of all of the non-excluded tags (hence tags that are either included or set to neutral). On the right part of this view, a list of included, excluded and neutral tags is provided, showing tags as bar elements colored with their corresponding category color, with the aforementioned slight saturation and values offset (to also act as a legend for colors of tag elements of the various coordinated views). A ranking of these included, excluded and neutral tags by various attributes is also possible: unlike the previously described tags ranking, this also takes into account filters on tags, and thus ranks tags by their attribute based only on the filtered game results with the corresponding associated tag.

3.6. TAGS SIMILARITY SCATTERPLOTS & CHORD DIAGRAM

On the top right of the user interface, three possible visualizations for the similarity of tags can be accessed: a t-SNE scatterplot, an MDS scatterplot, and a chord diagram. The similarity value and t-SNE and MDS coordinates of each pair of tags were computed offline as described in Section 2 of this paper. The two tags scatterplot show the result of the t-SNE and MDS dimensionality reduction techniques, and visualize each point, representing a tag, with the associated tag category's color. It also shows a list of the 10 most similar tags on hover of each point, shows excluded tags with a lower opacity with respect to non-excluded tags, and visualizes tag names underneath each point when a sufficiently small amount of tags is being visualized.

Both the t-SNE and the MDS visualizations can be zoomed in/out and implement semantic zooming, in which the names of each tag appear underneath the corresponding tag's point as the visualizations are zoomed in.

The chord diagram visualization was inspired by the "TagReel" system proposed by Bae et al. [3], in which a "non-ribbon" version of traditional chord diagram visualizations is used in a radial chart. Tags are positioned around a circle as points, grouped by category (with outwards going names) and arcs are used to link tags that have a similarity value greater than 50%. The arcs' color is determined by the category color of the associated pair's "dominant" tag, defined as the tag with the highest number of associated games. The arc's opacity depends on the similarity value between the 2 tags: the higher the similarity, the higher the arc's opacity. Tags around the chord diagram are grouped by their category, hence by color, while Bae et al. [3] group tags by both color and point's spacing around the circle. As for the interaction method, Bae et al. don't propose a way of interacting with the visualization, but limit their solution to showing a fixed visualization of arcs linking points. The solution proposed in this paper instead allows the user to hover over points to show a list of similar tags in the form of a tooltip, and also to click and hold onto one of the points (or text labels) of the visualization to highlight the tag's point itself, all of its outgoing arcs and all of the linked points, by hiding every other point and arc, to allow for a local exploration of similar tags.

3.7. TAGS SCATTERPLOT / BUBBLE CHART

On the center right part of the system's user interface, a tags scatterplot can be visualized, showing points corresponding to tags with the respective category's color. The chart highlights non-excluded tags in a similar way to the similarity scatterplots, as previously described.

This scatterplot visualization allows users to choose which tag attribute to plot on both the X and Y axis (number of games, total revenue, average revenue, total copies sold, average copies sold, average review rating or average price). The values for the coordinates of each point in the chart are computed by only considering games that meet the current filter criterias, and therefore by considering only the tags associated with these games when computing the average and total values of these attributes. Additionally, the user can choose to encode a third tag attribute with the radius of each point, effectively transforming the tags scatterplot into a bubble chart. The scale of each axis can be switched independently between linear or logarithmic.

On hover over the chart area, the X and Y coordinates are shown right below the cursor, with the corresponding X and Y attribute values updated on mouse movements, while on hover over tag points, the values of each of the tag's attributes is shown using a dedicated tooltip.

The user can ultimately also interact with the graph by using brushing: on click and drag onto the scatterplot, a rectangular selection area will be shown, and at the end of the interaction, tags inside the defined area will be included and/or excluded based on whether the user used the right or left mouse button (tags are included or excluded with the same modalities described for the tag's ranking visualization, and produce the same results on filtered games).

3.8. GENERAL GAMES INFO

The general games info view, on the bottom left of the system's UI, shows numerical information about the currently included games in the filter results, in three separate sections. The first section shows the percentage and number of games results being shown based on the current filters and with respect to the total number of games. The second section shows the total revenue and total copies sold of all the games included in the filters, with also their corresponding max values (of the included game item with the highest estimated revenue and the highest estimated number of copies sold respectively) as well as the variation, in percentage, with respect to the same values for all the games in the database. The third and last section of this view shows the average revenue, average copies sold, average review rating and average price of the included game results, with the associated increase/decrease percentage with respect to all the games in the system's database.

3.9. GAME RESULTS

The games results section at the bottom of the system's user interface presents all of the games which meet the filter criterias defined by interacting with the other coordinated visualizations, in a tabular form.

The table shows the main game attributes (name, release date, tags, price, review rating, estimated number of copies sold and estimated revenue). By hovering over a game's name cell, all of the game's information available in the dataset are shown (languages, content ratings, player modes, developers, publishers, ecc…). By clicking onto a game item's name, the Steam storefront page for that specific item is opened in a new tab (redirecting to Steam's website).

The visualization also allows to sort the game results by the attributes corresponding to the table's columns, in both ascending or descending order, by clicking onto the header row's cells. A particular type of sorting method is that of the "Tags" column of the table: it allows to order games by "tag relevance", hence the relevance of the game's tag to the active filters on tags (it uses the ranking of the tags associated to each game to determine the most relevant game items based on the included tags in the filters).

Ultimately, this visualization allows to exclude single game items from the filter results (and thus from all of the other visualizations of the system) by clicking onto the last cell of the corresponding game entry in the table.

3.10. TAGS TIME SERIES

The tags time series visualization, at the bottom right corner of the system's UI, allows users to visualize the value of a selected tag attribute over time. It plots lines, corresponding to tags, based on the selected time intervals ("day", "month", "quarter", "semester" or "year"): the chosen time interval will determine the size of the buckets corresponding to the plotted lines' points. Additionally, other than choosing which attribute to plot on the Y axis of the graph, the user can decide to add an additional line for each of the already existing tag lines (which will be shown as a dashed line), plotting a second attribute chosen from the same list of attributes, and making a second (dashed) Y axis appear on the right. Hovering over the chart shows information about the X, Y1 and Y2 cursor position's coordinates (hence the related axes' attribute values and time intervals).

4. RELATED WORKS

This section will present related works for existing market analysis solutions in the information visualization field.

As already mentioned in Section 1 of this paper, not a lot of recent works exist in the literature for business data visual analytics solutions, as also noted by Roberts et al. [5] in their extensive survey centered around visualizations for business data. This can be attributed to the lack of publicly available data in the business fields, and thus in the lack of research material to work with for this kind of visual analytics solutions. Businesses and organizations operating in specific markets often rely on competitive advantages over other organizations, which leads to the need to keep their data private and to conduct business data analysis internally, making sure not to let this data leak out of the organization to avoid adverse economic, legal and social effects [17].

This lack of data and transparency obviously negatively impacts the scientific and research landscape for the business information visualization field. The slow down of this research field have been especially detrimental in the most recent years, in which major transformations took place across different industries, influencing the way organizations conduct their business, most notably due to the Covid-19 pandemic of 2020, which significantly accelerated the digital transformation of business (among other aspects of the human life) [9]. While the industry field had to react quickly to these rapid shifts of paradigms and customer behaviors, the scientific field, when it comes to business related visual analytics solutions, lagged behind.

The work tried to identify similar research efforts and solutions, but the search was not very successful, with only a few solutions being close enough to the proposed work, most of them also being works published 6 years ago, focusing on traditional business practices centered around geographical distributions of shops and physical sales location [2] rather than online markets and digital commerce platforms, or sometimes on specific markets rather than broad consumer markets and large marketplaces of various types of goods and services [5]. Nonetheless, the following will briefly describe the existing literature works which were the closest to the ideas and aims of the proposed solution, and will then highlight the differences with respect to the latter.

4.1. FOOD & BEVERAGE CASE STUDY

We will start with a case study on the "Food & Beverage" industry for visual analytics solutions applied to market research, by Mann et al. [4]. This is the most recent related work, published less than 2 years ago, and its topics and analyzed problems were also very close to what the proposed solution aims to address. The work did an excellent job at capturing the importance of market research for companies and organizations as globalization becomes more and more relevant over the years, and also managed to explain how visual analytics solutions applied to market research can help in easing the analysis, the understanding and the accessibility of business data and information.

The paper work then, however, proceeds in applying visual analytics solutions rather poorly, with simple static charts and infographics rather than interactive data visualizations and coordinated views. Furthermore, the focus of the case study, the "food and beverage" industry, is diametrically opposed to the industries and markets considered by the solution proposed in this paper: while the former focuses on selling physical goods and thus on the analysis of geographical data on sales, the latter (the solution presented in this paper) focuses on an online market for digital products (games), and can be extended to e-commerce and online marketplace platforms in general, as well as to digital goods and services marketplaces. The differences in markets, business procedures and consumer behaviors of these two different subject matters leads to big differences in the way the market analysis should be carried out, and thus in the visual analytics and information visualization solutions that can be used for this task.

In the proposed solution, the focus is on more recent developments of industries and businesses, with attention put into the way modern organizations conduct their business, and thus a focus on e-commerce platforms, large online marketplaces, and digital goods and services consumer markets.

The integration of the information visualization field for the proposed solution is also stronger than the one proposed by Mann et al., with various interactive and coordinated views rather than static infographics that cannot well adapt to the user needs and to the development and changes of the underlying data.

4.2. VISUAL ANALYTICS FOR COMPETITIVE INTELLIGENCE

The second analyzed paper describes "MarketAnalyzer", an interactive visual analytics system for analyzing competitive advantage using point of sale data, proposed by Ko et al. [2]. It is a visual analytics system designed for businesses' "competitive intelligence", hence the exploration, analysis and prediction of the market share of businesses operating in a specific market [16]. Competitive Intelligence aims at forecasting changes in a business's or company's market to ultimately expand the company's market share by analyzing competitor companies, evaluating the potential of new business strategies, and identifying market risks.

The MarketAnalyzer system provides visual analytics tools to explore current sales volumes, trends and temporal market share growth rates within a market. It allows to analyze "points of sale" (POS) data, i.e. the information about customers, products and physical/online stores that is collected during the customer sales transactions.

The interactive, coordinated aspect of the MarketAnalyzer system aimed at analyzing market data is very similar to what the proposed solution implements. The system proposed by Ko et al. employs pixel-oriented matrices and stacked bar chart visualizations to compare sales, trends, and growth of a main company and a selected secondary company over a certain time interval. It also employs other minor visualizations, such as geographical maps for physical stores data and time series line charts. The MarketAnalyzer system is therefore centered around "products", "stores" and "companies", and is very effective for visualizing and analyzing data for markets where these three entities are all equally relevant for determining trends, growth rates and characteristics of the underlying market. When "companies" and "stores" are significantly less important than single "products", however, the MarketAnalyzer system fails: these types of markets are represented by e-commerces, software or app stores and digital goods and services online marketplaces, which are instead the focus of the solution proposed in this paper. The MarketAnalyzer solution was proposed more than 10 years ago, and relies on traditional business and market analysis practices, centered around the sales of physical goods, with no mention of online marketplaces and e-commerce platforms, not yet established at the time of the publication.

In large online marketplaces customers don't shop by physically moving through stores of a certain brand, but by navigating online platforms where various brands and companies might be selling the same product, and therefore where the single products become the main focus of both customers and businesses. In this case, the process of customers' product discovery involves the use of interactive search filters, to limit the scope of the search, to categorize or group products together by means of labels and tags. This does not happen in physical markets, where customers discover products by navigating physical and confined spaces, in which each product's visibility and accessibility is determined by their physical placement in the navigated space, dictated by companies or store owners (e.g. supermarkets, shopping center stores, ecc...).

This change in the product discovery process leads to major changes also in companies' business processes and models.

When dealing with online markets, single competitor companies and (most notably) stores lose their relevance, in favor of single products, which need to be associated with a wide variety of data to represent their features and characteristics (e.g. product descriptions, images, categories, tags, ecc...) to ease the discovery process of customers. This large amount of data for every single product can then be leveraged by businesses for market analysis, which therefore takes the form of a qualitative analysis other than just a simple quantitative analysis (where qualities, hence categories, features and characteristics of single products, become as important as quantities, hence sales and revenues). The solution proposed in this paper does in fact leverage the large amounts of data available for each product for market analysis, and in particular exploits the associated tags of each product to let users discover business insights.

4.3. OTHER RELATED WORKS

This paper already discussed the “TagReel” system by Bae et al. [3] (Section 3.6 of this paper), with a focus on the technical aspect of the visualization proposed in the paper, and the two papers by Mann et al. [4] and Ko et al. [2] (Sections 4.1 and 4.2 of this paper), with a focus on methods and topics of the proposed solutions instead. This next section of the paper will continue describing the panorama of related works centered around business-oriented visualization systems, mainly through the analysis of survey research works on this topic.

The most comprehensive survey for business data visualizations was published by Roberts et al. [5] at the end of 2018, and presents a distinction of these kinds of visualizations into four categories: “Business Intelligence”, “Business Ecosystem”, “Customer Centric” and “Financial Visualizations”. For each of these four categories, a wide variety of existing visual analytics solutions are presented, with some case studies extensively discussed throughout the survey. Out of all of the solutions presented in this survey, few are centered around visual analytics solutions for market analysis (one of them being also the already discussed “MarketAnalyzer” solution by Ko et al. [2]), and none are centered around online marketplaces, e-commerces and digital stores. This is surprising if we consider that, although the survey only includes visual analytics solutions published before 2019 (with the oldest being 1997 and 1999 works by Wright and by Wattenberg), the development of digital commerce as we know it today was already well established as of the 2000s and the 2010s [18], after the “World Wide Web” made content sharing over the internet possible in the 1990s and after the launch of the first online marketplaces platforms that still remain relevant as of today (Amazon.com and eBay.com in 1995 and Alibaba.com in 1999 as massive online marketplaces, Steam in 2003 and online games and digital products stores by other major digital players, like Google, PlayStation and Microsoft, subscriptions services like Netflix in 1997, but also big physical stores launching their e-commerce equivalents, like Walmart in 2000, and ultimately the first platforms that provide services to ease the launch of e-commerces themselves, like Shopify in 2006).

The “Business intelligence” category of the solutions analyzed in the survey by Roberts et al. is the category in which the solution proposed in this paper would fall into. Business Intelligence (the strategies and technologies used by enterprises for the data analysis and management of business information) translated to the digital markets space is often conducted through the analysis of data and visualizations directly provided by the integrated tools of the online marketplaces and e-commerce platforms, and is often relegated to “Internal Business Intelligence” (analysis of the internal data of a single company) rather than “External Business Intelligence” (analysis of data of a company with respect to other companies, oftentimes competitors).

The business intelligence tools associated to marketplace platforms oftentimes don't provide a detailed breakdown of all the companies' sales data and private information, but rather just detailed information about individual companies (e.g. sales, customer profiles and general POS data): this is done to preserve the privacy of each business and to allow marketplaces to remain a fair competitive environment for all the entities operating on their platform. As already discussed in the introduction for Section 4 of this paper, this missing disclosure of private business-oriented data hinders both research developments and the actual implementation of external visual analytics solutions for comprehensive market research, in favor of integrated tools in online marketplaces platforms for companies business intelligence, limited to the display of one company's private data, without providing a global and comprehensive view of the entire marketplace as a whole. An exception to the general unavailability of global market data is represented by financial markets, in which global data about sales, volumes, trading entities, and general information for each product is entirely available to all of its players (investors, funds, banks and corporations). A separate survey by Ko & Cho et al. [7] has been published for financial data visualization, but is outside the scope of the proposed work. It is not a surprise, therefore, that almost the entirety of the visual analytics solutions for market analysis and market research focuses on the financial market rather than on specific markets.

Another analyzed survey, this time centered around commercial visual analytics systems and published by Behrisch et al. [1] in 2018, describes the evolution of commercial Visual Analytics systems, focusing on the analysis and evaluation of their features, performance, and usability from the point of view of their users, classified into three main groups: upper management, domain experts, and data analysts/engineers. The paper evaluates ten case-study visual analytics solutions based on criteria like data handling, automatic analysis, complex data types, visualization, and user guidance, and emphasizes the need for better exploratory analysis support and user-guided exploration. This survey work is a general analysis of commercial visual analytics systems and doesn’t focus on business data visualizations, but it still manages to capture the importance of visualizations for big data, which are at the center of market analysis for businesses and organizations. The described methodologies for evaluating commercial visual analytics systems, in particular for the data analysts user group, were taken into account for the implementation of the solution proposed in this paper, in particular the need for extensibility, interactivity, and data handling when conducting confirmatory, hypothesis-driven and exploratory analyses.

5. USE CASES & INSIGHTS

This section will present the proposed solution's intended users together with a list of general possible use cases, and then will dig deeper into examples of insights obtained by interacting with the implemented tag-oriented business data visual analytics system.

5.1. INTENDED USERS AND USE CASES

The following is a list of intended users and associated general use cases for the proposed system solution:

- Large-size video game studios developing games in-house and publishing them in a wide variety of gaming platforms: in this case, the proposed solution can be used for conducting an initial market analysis phase on the Steam market (analyze competitors, assessing product-market fit, visualizing trends and growths of the industry as a whole and of its various sectors);
- Small independent game studios in the process of self publishing a game on Steam (developing a game and self publishing it as a small independent studio can take years, and it is crucial to be able to make estimates about sales and thus understand whether the game will grant a way for the studio to remain financially sustainable and eventually cover the expenses of the development process at its release);
- Video game publishers (hence companies funding game studios, and thus covering the initial development costs until the release of the game, to then first recoup the funding costs and subsequently earning a percentage fee of any sales of the game) looking for a game product to publish, with the ultimate goal of profiting from the launch of the game: in this case, the proposed solution will provide a way for publishing companies to evaluate market viability of game product ideas and prototypes pitched by video game software houses, development companies or teams;
- Gaming software houses and developers looking for game publishers to work with: in this case, the proposed solution can be used to conduct a market analysis first to estimate the size of the potential market, the potential revenue, and find similar existing products which were financially successful, and ultimately use these data as ready-made market research to pitch the game idea or product prototype to publishers.
- More use cases for video game software developers and publishers, with some examples briefly mentioned here:
  - Discover genre and sub-genre "niches", with a low product supply but high demand (to identify the features of a potential product to fit these niches);
  - Evaluate the current and recent market viability of a product that a company has been developing (as big video game software companies might take several years to develop a game, and it may be necessary to reassess the potential return on investment of the product in development and decide whether to reallocate resources towards or away of it);
  - Identify trends to predict the desktop gaming market direction of the near future (and potentially head-start the development of a product in that direction);
  - Help identify product, price, placement and promotion (the four 'P's of marketing) by researching the market for successful genres, average prices of products in said genres, the main features of those products and already existing products presentation and placement for their potential customers and users.

The proposed system solution was aimed towards these types of users and potential use cases, but it was also developed to maintain a high versatility with its tasks and goals. This was accomplished by allowing users to choose different data features to visualize in different ways for the various views of the system (e.g. change plotted features on axes for the various system's tag visualizations, change ranking and sorting criterias of both games and tags visualizations, allow grouping of data in rankings and charts, switch between multiple tag similarity views, excluding single games and tags results from visualizations, highlight of data points in the various visualizations in a coordinated way).

5.2. USE CASES & INSIGHTS EXAMPLES

The first insight example will simulate the use of the implemented system by a small independent game studio, made of a small number of developers (from one to five), looking forward to self publish a game on the Steam platform. The goal is to decide which game the studio should develop next. In doing so, analyzing the performance of existing games of a certain genre and subgenre which are similar to games the studio can afford to develop is paramount for the success of the studio and its financial sustainability.

The development team can use the system by first restricting their search to games with an "English" language (they cannot afford to pay for a translation of their game) and with a "Single Player" player mode (as they cannot afford to pay for servers and services to enable an online multiplayer feature of the game). The studio also decides to restrict their search to recent games, hence games released within the last two years. By using the interactive visualizations and setting these filters, the studio discovers that there are about 26.500 games matching their search filter criterias, out of the 73.000+ games published on Steam. The studio also finds out that 52% of these games sold 0 copies, while the remaining games seem to have sold around 100-999 copies, with only 20 games selling more than 1 million copies. The users therefore decide to restrict their search even more, to games that sold between 1000 and one million copies, since the studio is confident enough to know that the quality of their product will allow them to sell at least 1000 copies of their game, but does not anticipate to sell more than 1 million copies, as these numbers are usually relegated ot big studios spending millions of dollars on the development and promotion of their games. They therefore narrowed their search results to 2.933 games. For these games, the next step is to look at the tags ranking by average revenue and thus find out about the "Sub-Genres" tags with the highest average revenue. By interacting with the tags ranking visualization, the tags grouping and tags inclusion/exclusion feature, the team identifies the top 50 tags by copies sold, including them in the search filters while excluding all other tags (also excluding all "Genres" and "Features" tags). By looking at the information about these 5 tags, the users find out that some of these tags are associated with less than 10 games, and thus the average revenue of said tags gets highly skewed towards the best seller game. The team therefore decides to use the system to rank tags based on their number of games, and then use the system to exclude from their search filters 11 tags associated with less than 10 games. The team now decides to rank the selected tags (which are all "Sub-Genres" for games with a high average revenue) by total copies sold, to determine the Sub-Genres with the widest possible audience, to be sure to choose a genre for their game with a wide appeal to players.

By scrolling through these sorted tags, they are looking for candidate genres to start developing their next game, by being careful not to choose a sub-genre which is outside of the scope of competences and resources the team has.

The team ultimately decides to go for the 10th tag of the list, the "City Builder" sub-genre, and, by looking at the tags stats, find out that, for games with an english language, a "single player" player mode, and released after January of 2022, this tag sold an average of 34.000 copies, has an high average revenue of 617.000 dollars, which is 88% more than the average revenue of all games of the Steam platform, has an average game item price of 17 dollars, and that its games have an average review rating of 82% positive reviews.

A second insight example may be for a publisher company specialized in Virtual Reality (VR) games, that received a pitch-deck from an independent game studio looking for funding. The pitched game idea is about a VR “Online Co-Op” shooter game, of which the publishing company played a demo for, and which showed promising results. In order to evaluate the market viability for this kind of game and also to find out about existing similar competitor games. The publishing company starts by restricting their search to game items tagged “VR” and “Shooter”, thanks to the tags filter of the system. Filters on the “Multiplayer”, “Co-Op” and “Online Co-Op” player modes were also included. A total of 159 games resulted from this search, of which the majority sold less than 1000 copies. The search is then restricted to games that sold at least 1000 copies, resulting in a total total of 92 existing games with these characteristics on the Steam platform. The next step is to look at the list of game results for these filters, and exclude a few games that seem to be related to famous intellectual properties and some non VR-exclusive games. The average revenue for the resulting games is 660.000 dollars, with also an average copies sold of 54.000 copies: being specialized in VR games, the publishing studio knows that these are good numbers for VR games (which usually have smaller numbers compared to desktop games since they require additional hardware, namely a VR headset, to be played), and therefore decides to propose a publishing deal to the developers of the game.

6. CONCLUSIONS AND FUTURE WORKS

This work aimed at presenting a solution for the market analysis of massive online consumer marketplaces using a tag-oriented approach paired with visual analytics solutions. The presented solution was developed for the “Steam” platform, an online marketplace for desktop video games, but can easily be generalized to other consumer online marketplaces, e-commerces and digital products and services online stores. The advantage of using a “tag-oriented” market analysis approach is the possibility to analyze product’s business data (e.g. about sales, customers or revenue) even in presence of very dishomogeneous collections of physical/digital products or services, by means of analyzing data about single tags instead of single products, and thus discover insights about overlapping subsets of items/products that might never be discovered by looking at the single products or product category themselves.

The analyzed online marketplace (the “Steam” platform) offered publicly available data to be directly used in the analysis, while private data (sales data) could only be computed using estimates of sales and revenues of the various products available on the platform. It is important to note that using official business-related data about marketplace platforms instead of estimates might certainly lead to more precise results and insights. To the products of the Steam platform, sets of tags were assigned by the respective distributors, and these tags were later refined by the customer base of the platform by proposing changes and additions. Tags were used in the proposed solution to compute new data and offer additional insights on the overall structure, shares and trends of the gaming market on desktop platforms. The underlying structure of the tags associated with game items is that of a simple set, with no defined mutual relationships and only a simple classification of tags into three main categories as additional data (tag categories presented in the paper, hence “Genres”, “Sub-Genres” and “Features”). Possible future works might be focused on extending the subject of the proposed solution, hence the tag-oriented, visualization assisted market analysis, to hierarchical tag-based structures for the underlying products: in the analyzed case (Steam online platform's products), in fact, while a classification of tags in three categories was possible, no tag hierarchy could be defined. Introducing hierarchical relationships between tags might help discovering different and possibly richer insights about the analyzed market, which may ultimately lead to more informed business decisions.

REFERENCES

[1] Behrisch, Michael & Streeb, Dirk & Stoffel, Florian & Seebacher, Daniel & Matejek, Brian & Weber, Stefan & Mittelstädt, Sebastian & Pfister, Hanspeter & Keim, Daniel. (2018). Commercial Visual Analytics Systems-Advances in the Big Data Analytics Field. IEEE transactions on visualization and computer graphics. PP. 10.1109/TVCG.2018.2859973.

[2] Ko, Sungahn & Maciejewski, Ross & Jang, Y. & Ebert, David. (2012). MarketAnalyzer: An Interactive Visual Analytics System for Analyzing Competitive Advantage Using Point of Sale Data. Computer Graphics Forum. 31. 1245-1254. 10.1111/j.1467-8659.2012.03117.x.

[3] J. Bae and K. Lee, "TagReel: A Visualization of Tag Relations among User Interests in the Social Tagging System," 2009 Sixth International Conference on Computer Graphics, Imaging and Visualization, Tianjin, China, 2009, pp. 437-442, doi: 10.1109/CGIV.2009.69.

[4] C. K. Mann and P. Subramanian, "Visual Analytics for Market Research: A Case Study of F&B Industry," 2022 IEEE 2nd International Conference on Mobile Networks and Wireless Communications (ICMNWC), Tumkur, Karnataka, India, 2022, pp. 1-5, doi: 10.1109/ICMNWC56175.2022.10031808.

[5] Roberts, Richard C. and Robert S. Laramee. “Visualising Business Data: A Survey.” Inf. 9 (2018): 285.

[6] Martin Bustos - Steam Games Dataset. www.kaggle.com/datasets/fronkongames/steam-games-dataset

[7] Ko, S.; Cho, I.; Afzal, S.; Yau, C.; Chae, J.; Malik, A.; Beck, K.; Jang, Y.; Ribarsky, W.; Ebert, D.S. A Survey on Visual Analysis Approaches for Financial Data. Comput. Graph. Forum 2016, 35, 599–617

[8] ColorBrewer - Qualitative & Sequential Color scheme for standard system visualizations’ color encodings. colorbrewer2.org/#type=qualitative&scheme=Set1&n=4

[9] Kutnjak, Ana. (2021). Covid-19 Accelerates Digital Transformation in Industries: Challenges, Issues, Barriers and Problems in Transformation. IEEE Access. PP. 1-1. 10.1109/ACCESS.2021.3084801.

[10] Zion Market Research - PC Games Market By Genre (Role-Playing, Action, Racing, Adventure, Strategy, Sport, Fighting, And Others), By Type (Physical, Digital, And Online Microtransaction), And By Region - Global And Regional Industry Overview, Market Intelligence, Comprehensive Analysis, Historical Data, And Forecasts 2023 – 2030 www.zionmarketresearch.com/report/pc-games-market

[11] DemandSage - Steam Statistics For 2024 (Users, Popular Games & Market) www.demandsage.com/steam-statistics

[12] Jake Birkett, Ryan Clark, Mike Boxleiter - How to estimate how many sales a Steam game has made greyaliengames.com/blog/how-to-estimate-how-many-sales-a-steam-game-has-made/

[13] Steamworks Documentation - Tags partner.steamgames.com/doc/store/tags

[14] Steam Announcement: Closing Greenlight Today, Steam Direct Launches June 13 steamcommunity.com/games/593110/announcements/detail/1265922321514182595

[15] Rune Skovbo Johansen - Steam Ratings Percentages Breakdown twitter.com/runevision/status/1307399794000887810

[16] KAHANER L.: Competitive Intelligence: How to Gather, Analyze and Use Information to Move Your Business to the Top. Touchstone Press, New York, U.S.A., 1998.

[17] Juma'h, Ahmad & Alnsour, Yazan. (2020). The Effect of Data Breaches on Company Performance. International Journal of Accounting and Information Management. 28. 10.1108/IJAIM-01-2019-0006.

[18] Tian, Yan & Stewart, Concetta. (2007). History of E-Commerce. 10.4018/9781599049434.ch001.