A3: Exploratory Data Analysis With Tableau

CSE 412: A3 Exploratory Data Analysis With Tableau

Jaiden Atterbury, jatter41@uw.edu, Section AA

Part 1

Data Selection

Board Game Data

The Dataset

	game_id	game_type	maxplayers	maxplaytime	minage	minplayers	minplaytime	name	playingtime	yearpublished	category	mechanic	average_rating	users_rated	average_complexity
1	1	boardgame	5	240	14	3	240	Die Macher	240	1986	Economic,Negotiation,Political	Area Control / Area Influence,Auction/Bidding,Dice Rolling,Hand Management,Simultaneous Action Selection	7.66508	4498	4.3477
2	2	boardgame	5	90	12	3	30	Dragonmaster	30	1981	Card Game,Fantasy	Trick-taking	6.60815	478	1.9423
3	3	boardgame	4	120	10	2	30	Samurai	60	1998	Abstract Strategy,Medieval	Area Control / Area Influence,Hand Management,Set Collection,Tile Placement	7.44119	12019	2.5085
4	4	boardgame	4	60	12	2	60	Tal der Könige	60	1992	Ancient	Action Point Allowance System,Area Control / Area Influence,Auction/Bidding,Set Collection	6.60675	314	2.6667
5	5	boardgame	6	90	12	3	90	Acquire	90	1964	Economic	Hand Management,Stock Holding,Tile Placement	7.3583	15195	2.5089
6	6	boardgame	6	240	12	2	240	Mare Mediterraneum	240	1989	Civilization,Nautical	Dice Rolling	6.52534	73	3
7	7	boardgame	2	20	8	2	20	Cathedral	20	1978	Abstract Strategy	Area Enclosure,Pattern Building,Pattern Recognition,Tile Placement	6.50534	2751	1.8217
8	8	boardgame	5	120	12	2	120	Lords of Creation	120	1993	Civilization,Fantasy	Modular Board	6.14538	186	2.4
9	9	boardgame	4	90	13	2	90	El Caballero	90	1998	Exploration	Area Control / Area Influence,Tile Placement	6.51776	1263	3.1958
10	10	boardgame	6	60	10	2	60	Elfenland	60	1998	Fantasy,Travel	Card Drafting,Hand Management,Point to Point Movement,Route/Network Building	6.74996	6729	2.1649
11	11	boardgame	7	45	13	2	45	Bohnanza	45	1997	Card Game,Farming,Negotiation	Hand Management,Set Collection,Trading	7.06751	28354	1.6777
12	12	boardgame	5	60	12	2	45	Ra	60	1999	Ancient,Mythology	Auction/Bidding,Press Your Luck,Set Collection	7.47505	15378	2.356
13	13	boardgame	4	120	10	3	60	Catan	120	1995	Negotiation	Dice Rolling,Hand Management,Modular Board,Route/Network Building,Trading	7.26569	67655	2.3603
14	14	boardgame	4	25	10	3	25	Basari	25	1998	Negotiation	Roll / Spin and Move,Set Collection,Simultaneous Action Selection	6.78156	1476	1.8588
15	15	boardgame	6	90	12	2	90	Cosmic Encounter	90	1977	Bluffing,Negotiation,Science Fiction	Hand Management,Variable Player Powers	6.9347	3629	2.3708
16	16	boardgame	4	60	12	3	60	MarraCash	60	1996	Economic	Auction/Bidding	6.84341	877	2.1538
17	17	boardgame	2	5	10	2	5	Button Men	5	1999	Collectible Components,Dice,Fighting,Print & Play	Dice Rolling,Press Your Luck	6.3087	724	1.5493
18	18	boardgame	8	120	12	2	45	RoboRally	120	1994	Miniatures,Racing,Science Fiction	Action / Movement Programming,Grid Movement,Modular Board,Simultaneous Action Selection	7.15355	19371	2.434
19	19	boardgame	4	45	9	2	30	Wacky Wacky West	45	1991	American West,Bluffing,City Building	Tile Placement,Voting	6.31166	1459	1.8467
20	20	boardgame	4	90	12	2	90	Full Metal Planète	90	1988	Science Fiction	Action Point Allowance System	7.43592	602	3.1636
21	21	boardgame	7	0	12	1	0	Gateway to the Stars	0	1981	Civilization,Exploration,Science Fiction	NA	5.35714	28	3
22	22	boardgame	16	240	12	1	240	Magic Realm	240	1979	Adventure,Exploration,Fantasy	Action / Movement Programming,Modular Board,Rock-Paper-Scissors,Role Playing,Simultaneous Action Selection,Variable Player Powers	7.14384	1709	4.4985
23	23	boardgame	6	360	12	2	360	Divine Right	360	1979	Fantasy,Political,Wargame	Dice Rolling,Hex-and-Counter,Variable Phase Order	6.95654	483	3.1324
24	24	boardgame	6	240	12	2	240	Twilight Imperium	240	1997	Civilization,Negotiation,Political,Science Fiction,Space Exploration,Wargame	Dice Rolling,Hex-and-Counter,Modular Board,Tile Placement,Variable Player Powers,Voting	6.66812	667	3.4902
25	25	boardgame	6	200	12	2	200	Battlemist	200	1998	Exploration,Fantasy,Wargame	NA	5.93231	308	3.2105
26	26	boardgame	6	360	12	3	360	Age of Renaissance	360	1996	Civilization,Economic,Medieval,Renaissance	Area Movement,Auction/Bidding	7.0993	1944	3.8511
27	27	boardgame	6	340	12	2	340	Supremacy	340	1984	Economic,Political,Wargame	Commodity Speculation,Dice Rolling	5.57724	1201	3.1404
28	28	boardgame	8	180	12	2	60	Illuminati: Deluxe Edition	180	1987	Card Game,Humor,Negotiation,Political	Card Drafting,Dice Rolling,Route/Network Building,Tile Placement,Variable Player Powers	6.5289	4274	2.6489
29	29	boardgame	4	120	10	2	120	Terrain Vague	120	1993	Fighting,Humor	Action Point Allowance System,Modular Board	6.66639	61	3.2857
30	30	boardgame	4	90	10	1	90	Dark Tower	90	1981	Adventure,Electronic,Exploration,Fantasy,Fighting	Area Movement,Press Your Luck	6.6435	1052	1.8182
31	31	boardgame	5	90	10	2	90	Dark World	90	1992	Adventure,Exploration,Fantasy,Fighting,Miniatures	Dice Rolling,Grid Movement	5.18525	590	1.9592
32	32	boardgame	2	30	8	2	30	Bison	30	1975	Abstract Strategy,American West	NA	6.04137	182	1.6471
33	34	boardgame	8	180	12	1	180	Arkham Horror	180	1987	Adventure,Horror,Novel-based	Co-operative Play	6.56386	453	2.3111
34	36	boardgame	8	300	12	2	300	Federation & Empire	300	1986	Science Fiction,Wargame	Hex-and-Counter	6.23239	355	4.4098
35	37	boardgame	4	120	12	2	120	Dragon Masters	120	1991	Fantasy,Wargame	NA	5.69731	130	2.1
36	38	boardgame	4	30	10	2	30	Runes	30	1981	Deduction,Word Game	NA	6.06707	82	1.9
37	39	boardgame	4	60	12	2	60	Darkover	60	1979	Bluffing,Fantasy,Novel-based	NA	5.22865	96	2.4286
38	40	boardgame	4	120	12	2	120	Borderlands	120	1982	Bluffing,Civilization,Fantasy,Negotiation,Political	Area Movement,Trading	6.77517	254	2.7727
39	41	boardgame	4	30	9	2	30	Can't Stop	30	1980	Dice	Dice Rolling,Press Your Luck	6.85659	9095	1.1683
40	42	boardgame	4	90	12	2	90	Tigris & Euphrates	90	1997	Abstract Strategy,Ancient,Civilization,Territory Building	Area Control / Area Influence,Hand Management,Set Collection,Tile Placement	7.72379	20166	3.5309

Selection Process

Board games are a greatly forgotten about past time that have been linked with building memory, decision making, problem solving, and overall logical skills in kids, teens, and young adults. Board games also provide a way to decompress from the stressful realities of day-to-day life. In particular, the topic of board games is important to me, as it is one of the primary ways I used to spend time with my father when I was a child.

Before choosing this particular dataset, I was interested in doing this assignment on topics such as: the NFL, mountain climbing, natural disasters, disease prevelancies, car crash information, and even my own Spotify listening data. However, due to time restrictions, such as the time it'd take to obtain my Spotify listening data, as well as limitations in datasets I found on Kaggle, such as limited number of attributes/columns, I had to look elsewhere to find an interesting dataset that would be diverse enough to do a thorough and comprehensive exploratory data analysis.

What led me to choose this board game dataset specifically, aside from the reasons listed in the above paragraphs, is quite unique. What I mean by this is that I found this dataset with a bit of prior knowledge/previous inspiration. In particular, one of my classmates, who took STAT 302 last fall, mentioned something about a board game dataset that he had to use for one of his class projects. With that prior information, as well as board games playing a big role in my childhood, I was able to locate this dataset on GitHub without any issues. After taking a brief look at the columns I was able to decipher that this dataset would be diverse enough to create a well thought out and descriptive exploratory data analysis.

This dataset, which was scraped off of BoardGameGeek.com, contains information and user reviews on board games and board game expansions dating back centuries, although it is important to note that the exploratory data analysis itself will only focus on certain time periods of the board game dataset. This dataset is quite large, as it contains hundreds of thousands of rows all corresponding to different games. Due to my lack of expertise on the topics of games, I do not know if this list is comprehensive or not. The dataset that I am working with can be found on GitHub from this link. Despite a little bit of research, I was unable to find where the GitHub user originally got this data, for that reason I assumed that the given user scraped it off BoardGameGeeks' website themselves.

Due to issues with importing the dataset into my Tableau environment, I had to trim down the dataset from its original 90,400 rows. In order to trim down the dataset, I decided to only focus on the years 1919-2019, as well as focusing on games that have 25 or more reviews. Although the years designation was intentional; it allowed me to have a century worth of data, the 25 or more reviews was more or less an arbitrary value that I thought was a sufficient enough of reviews (and also one that is close enough to allow the Central Limit Theorem to apply when working with averages).

Investigative Questions

After briefly looking at the columns/attributes that the BoardGameGeeks dataset provides, I have three main research questions that I want to address. These questions include:

- How do the two types of board games compare? That is, what are the similarities and differences between board games, and board game expansions?
  - The difference between a board game and a board game expansion will be addressed in Part 2.

- How have certain aspects and features of board games changed over time? Has there been any noticeable changes, or have these aspects and features remained relatively constant?
  - As mentioned in the previous slide, the “over time” category will only focus on certain time periods. These time periods will be explained in Part 2.

- How do certain aspects and features of board games compare between the top categories and top mechanics of board games?
  - The definition of a game category and a game mechanic will be addressed in Part 2.

Part 2a

Exploratory Visual Analysis (Phase 1)

Phase 1 - Question 1: Variable Description

Q: What variables does the dataset contain?

A: As explained in the introduction slide, the board games dataset includes information about board games from a website called BoardGameGeek. In particular, there are three types of board game data provided in this dataset: identifiers, general board game information, and lastly user review data. The first of these types, as mentioned above, are called identifiers. The columns that I consider as identifiers are game_id, and name. These identifiers are what allow users to understand the specific game they are looking at. The second of the board game data categories is called general board game information. The columns that I consider as general board game information are game_type, maxplayers, maxplaytime, minage, minplayers, minplaytime, playingtime, yearpublished, category, and mechanic. These attributes allow users to understand basic information about the current game they are looking at. The last category appearing in the chosen board game dataset is labeled as user review data. The columns that I consider as user review data are average_rating, users_rated, and average_complexity. These attributes are special in the fact that they come from user reviews from the users of the BoardGameGeek website. Furthermore, these attributes give users information on games that is beyond the scope of general information that is intrinsically associated with each specific game. In total there are 15 columns and 19,344 rows. On the next slide, each variable will be described, as well as certain information about the values that each variable takes on.

Phase 1 Question 1 - Variable Description

- game_id: This variable represents the unique numerical identifier of a specific board game.
  - This variable is an integer that takes on a value from 1 to 224,316.
- game_type: This variable represents the particular type of the board game that each game represents.
  - This variable is a string that takes on the value of one of the two types of board games, which include: the normal game (boardgame) and an expansion of the normal game (boardgameexpansion).
- maxplayers: The maximum number of players allowed to play the game.
  - This variable is an integer that takes on a value from 0 to 999.
- minplayers: The minimum number of players required to play the game.
  - This variable is an integer that takes on a value from 0 to 10.
- maxplaytime: The expected maximum playing time of the game (in minutes).
  - This variable is an integer that takes on a value from 0 to 60,000.
- minplaytime: The expected minimum playing time of the game (in minutes).
  - This variable is an integer that takes on a value from 0 to 60,000.
- playingtime: The average playing time of the game (in minutes).
  - This variable is an integer that takes on a value from 0 to 60,000.
- minage: The recommended minimum age of a player allowed to play the game.
  - This variable is an integer that takes on a value from 0 to 120.

Phase 1 Question 1 - Variable Description

- name: The name of the board game.
  - This variable is a string that takes on one of 86,732 different game names.
- year_published: The year in which the game was published.
  - This variable is an integer that takes on a value from 1919 to 2019.
- category: The type(s) of game that the board game is.
  - This variable is a string that represents the type(s) of the game. These types include but are not limited to: Economic, Card Game, Fighting, etc.
- mechanic: The type(s) of ways players interact with the game and each other.
  - This variable is a string that represents the interaction type(s) of the game. These types include but are not limited to: Auction / Bidding, Trick-taking, Area Control, etc.
- users_rated: The number of users who rated the game.
  - This variable is an integer that takes on a value from 0 to 67,655.
- average_rating: The average rating of the game across all user ratings on a 0-10 scale.
  - This variable is a floating point value that takes on a value from 0.0-10.0.
- average_complexity: The average complexity of the game across all user ratings on a 0-5 scale.
  - This variable is a floating point value that takes on a value from 0.0-5.0.

Phase 1 - Question 2: Data Quality Issues

Q: Are there any notable data quality issues?

A: As expected with data scraped off of the internet, there were many problems with the quality of the data provided, below I will list some of the problems, as well as some of the problems that were fixed after filtering the data.

- One minor issue with the game_id column is that even before filtering the data, the unique identifier didn't go from 1 to 90,400 in ascending order. Instead, it skipped numbers randomly and went from 1 to 226,264. This issue was further exacerbated after filtering, as now there is no sense of order since specific rows were taken throughout the entire dataset.
    - This "problem" isn't a big deal since all an identifier column is supposed to do is uniquely describe each row in a dataset. Furthermore, the fact that the id column wasn't in ascending order from 1 to 90,400 shows that the original list of games wasn't comprehensive.

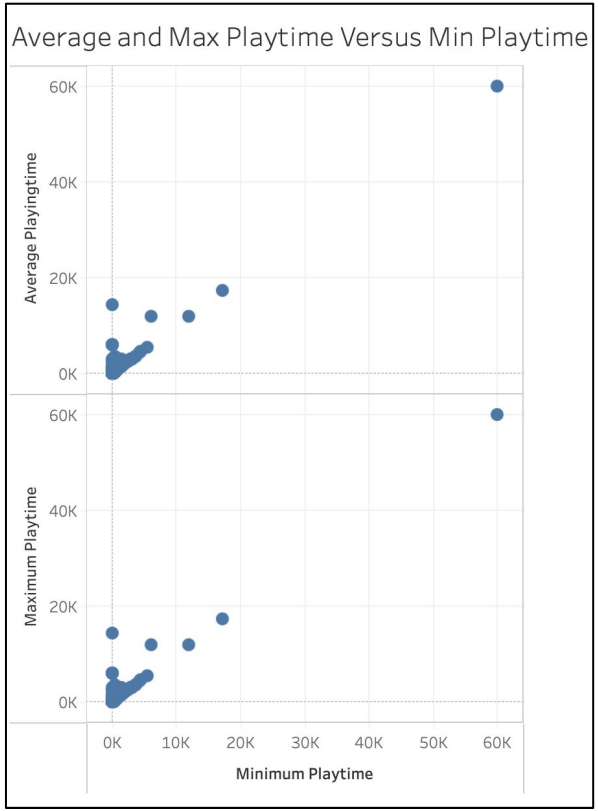
- One major issue that puts into question the validity of an entire attribute, is that the minage column has values such as: 120, 60, 45, etc. Since some of these values are an age that only a miniscule percentage of the world's population will get to, like 120, it follows that the majority of these high numbers make little to no sense to be considered as a "suggested minimum age for players." For example, the fact that *Star Wars: The Card Game* has a minage value of 60 makes no sense, as this game seems to be aimed towards kids and young adults in the first place.

- Another minor issue with the data is that the mechanic and category column are given as a comma separated list, which makes it difficult to work with on its own.
    - However, this problem can easily be dealt with in Tableau through the use of the SPLIT function. In the second phase of this analysis, the first entry in this comma separated list will be taken as the primary mechanic and primary category of the game.

Phase 1 - Question 2: Data Quality Issues

- The biggest issue that I noticed when doing my initial scan through the columns in the dataset was that the maxplaytime and playingtime variables have the exact same values. Furthermore, in many cases, the minplaytime value also shares the same value as maxplayingtime and playingtime. This relationship is apparent when looking at these scatterplots generated in Tableau.
    - After looking through the BoardGameGeek website, the only time variable that I could find corresponded to playingtime. This puts into question where the min and max playingtime variables came from in the first place.

- Two data quality issues that were fixed when the data was filtered in order to be ran in Tableau, were the year column and the average_ratings/average_complexity columns. First off, before the data was filtered based on the year range of 1919-2019, year values in the negatives and zero existed. Although some of these negative values were estimates of when ancient games such as *Tic-Tac-Toe* were invented, 0 was used as a default value for games that had no year associated to them on the game website. Also, by filtering the data based on games that have over 25 reviews, I was able to remove games that had skewed average_rating and average_complexity scores.



Phase 1 - Question 3: Data Distribution

Q: How is the data distributed?

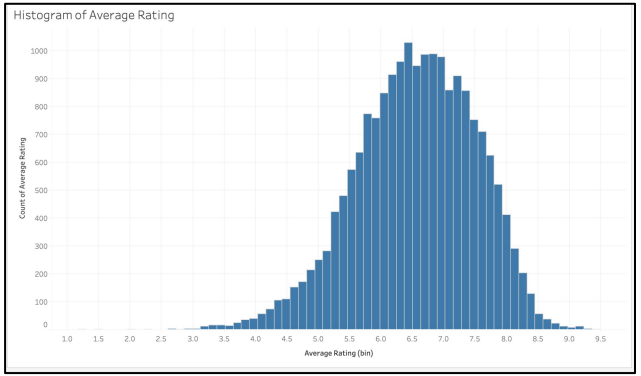
A: To answer this question, I split the columns/attributes into three types, in which all columns in a given category have similar distributions. The three categories include: no distribution associated, approximately normal, and ones that are hard to tell due to high outliers. The variables in these categories, as well as their associated distributions, are described below.

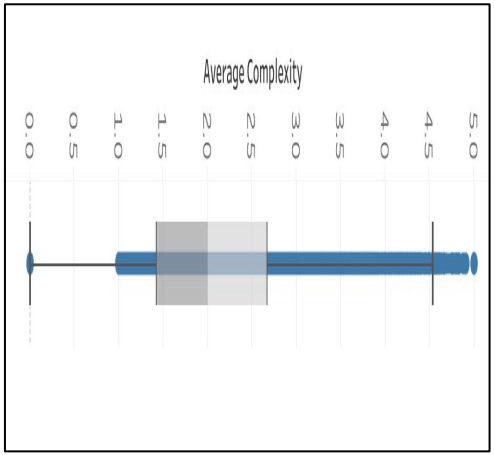
- No distribution associated:
  - The variables that don't have any associated distributions are either categorical variables or identifier variables. These variables that don't have an associated distribution are: game_id, game_type, name, year, mechanic, and category

- Approximately normally distributed:
  - The variables that are approximately normally distributed, unsurprisingly have to deal with averages. These variables are average_rating and average_complexity.
    - average_complexity is approximately normally distributed and centered around a mean value of around 6.7 (as seen in the top image on the right).

    - average_complexity is approximately normally distributed and centered around a mean value of around 2.1, with a minor right skew. (as seen on the bottom image on the right).

- Hard to tell:
  - The variables users_rated, minplayers, maxplayers, minage, and all of the playing time variables, all have distributions that are very difficult to discern due to the presence and amount of high outliers, as well as the fact that all of these variables are discrete. In particular, the histogram of each of these variables has a noticeable right skew, despite the fact that the majority of the data in these histograms are centered around a much lower value.





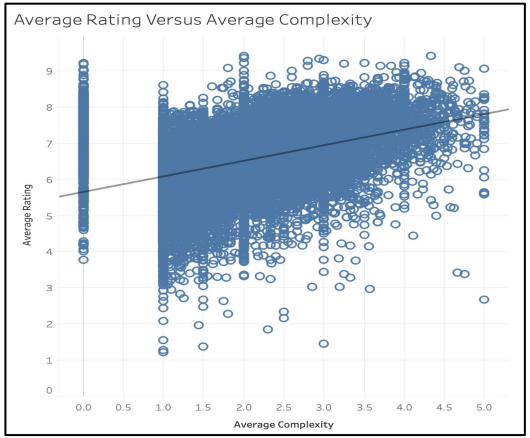
Phase 1 - Question 4: Surprising Relationships

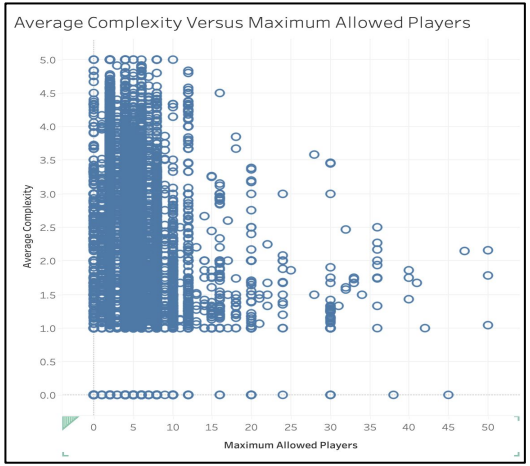
Q: Are there any surprising relationships among the variables?

A: To answer this question, I will look at the two types of relationships mentioned in the assignment specification, surprising relationships, as well as double checking relationships I expected. These two types of relationships are described below.

- Expected relationships:
  - Some expected relationships that I found when doing a quick perusing of the dataset were: a positive linear relationship between minage and average_complexity, as well as a positive linear relationship between playingtime and average_complexity. The most important relationship that I confirmed during this initial analysis of the dataset was the positive linear relationship between average_rating and average_complexity (as seen by the top image on the right). I expected this relationship to hold as the hobby of board games is very niche subject, and thus I expected these dedicated users to appreciate more complex games.

- Surprising relationships:
  - One surprising relationships that I found while confirming the expected relationships mentioned above was that average_rating does not increase linearly with users_rated. Instead, there is no relationship at all, until extreme values of users_rated where there appears to be a constant rating of around 7.5. The main surprising relationship that I found was that average_complexity does not increase with increasing maxplayers. Instead, it seems that average_complexity decreases with increasing maxplayers (after a certain threshold of around 10 maxplayers is reached). Initially, I was expecting the average_complexity to increase with maxplayers, but it seems as if the games with more players are more simple in nature.





Part 2b

Exploratory Visual Analysis (Phase 2)

Pages

Columns: Game Type | Average Rating (bin)

Rows: CNT(Average Rating)

Filters

Marks

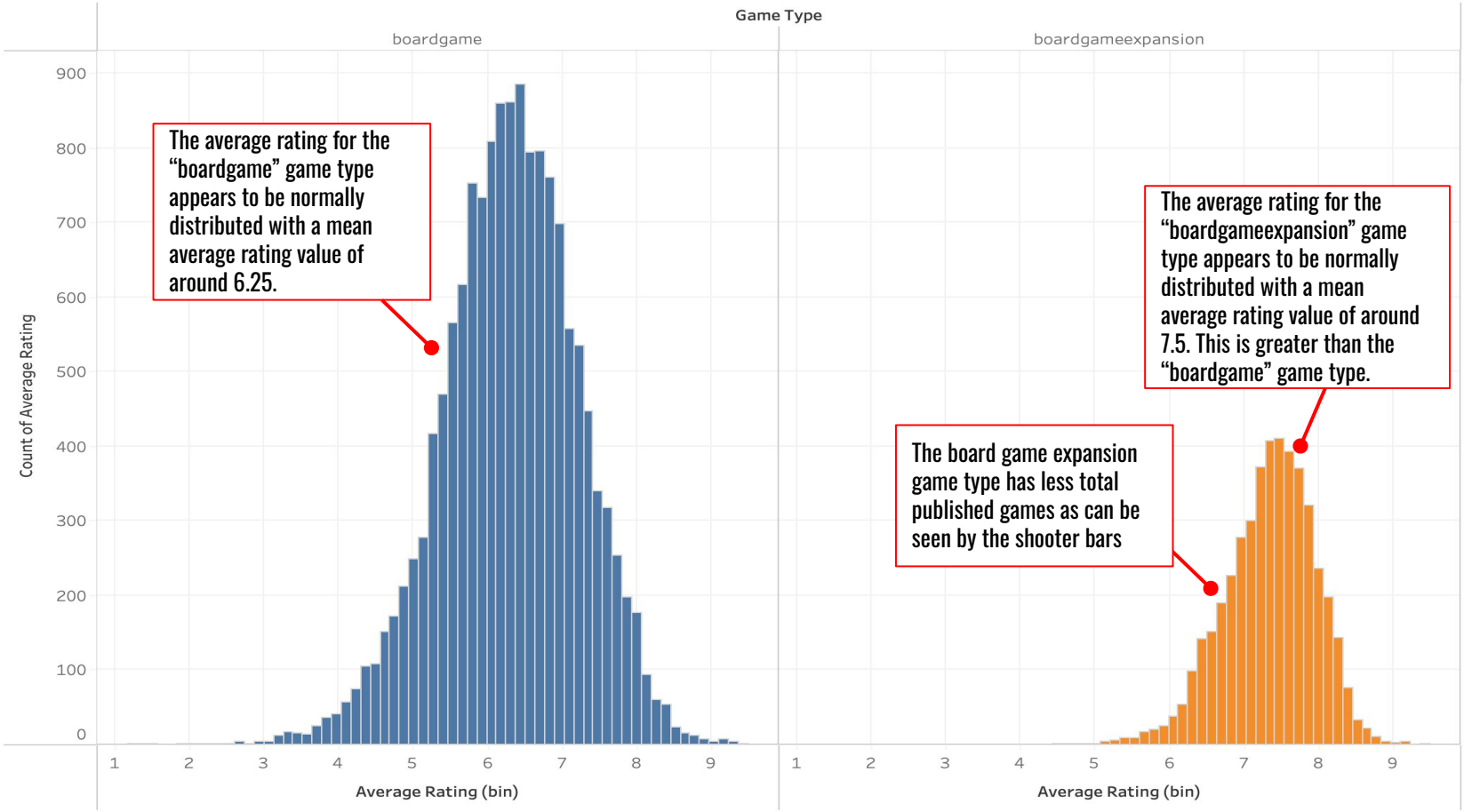
Automatic

Color | Size | Label | Detail | Tooltip

Game Type

Visualization 1

Histogram of Average Rating (By Game Type)

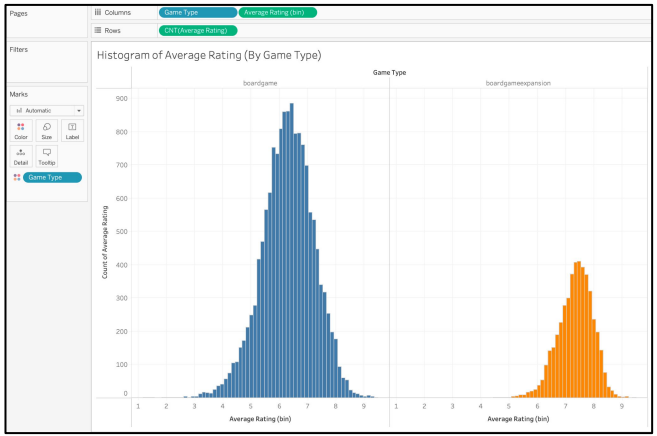


Visualization 1 Analysis

- **Caption:**
  - The pair of histograms above display the counts on the y-axis, for each of the different average rating (average_rating) bins displayed on the x-axis, for each board game type (game_type). The two types of board games are the original board game (boardgame) colored blue and expansions for board games (boardgameexpansion) colored orange.

- **Description:**
  - The goal of this visualization is to understand the distribution of the average rating of board games (average_rating) across the different board game types (game_type). In particular, this visualization answers question 1 from the above investigative questions list. As can be seen from the above histograms, the board game type called “boardgame” has a slight left skew in its distribution but can be considered approximately normally distributed with its center/mean around an average score of 6.25 with the associated count being around 870. Also, the board game type called “boardgameexpansion” has a slight left skew in its distribution but can be considered approximately normally distributed with its center/mean around an average score of 7.5 with the associated count being around 410. Both of the histograms have a relatively similar spread/standard deviation.

- **Insights/Takeaways:**
  - From the above histograms, and accompanying description, we can see that there have been a lot less expansions that have been created than there are actual board games. Furthermore, we can see that, on average, the average rating for expansions are much higher than those for the original board games. This makes sense intuitively, because in order to buy an expansion you must already like the game itself, and expansions are created to enhance the gameplay of the original game.



Pages

Visualization 2

Filters

Marks

Automatic

Color | Size | Label

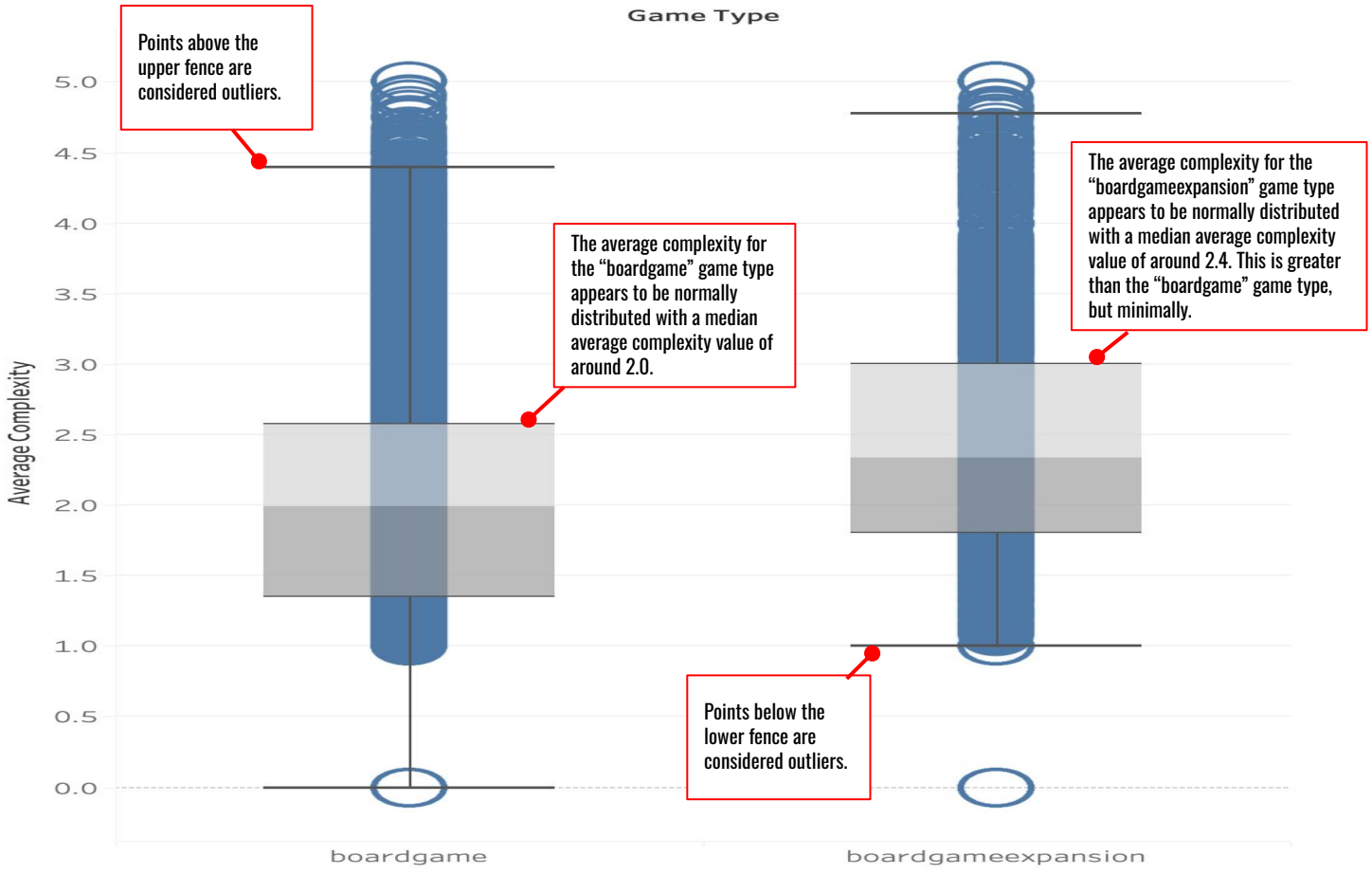
Detail | Tooltip | Shape

Game Type

Columns: Game Type

Rows: Average Complexity

Boxplot of Average Complexity (By Game Type)

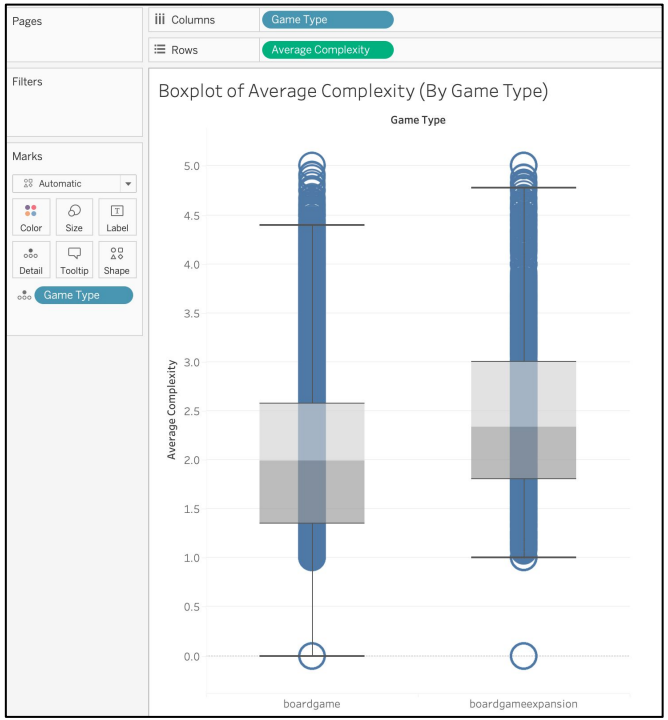


Visualization 2 Analysis

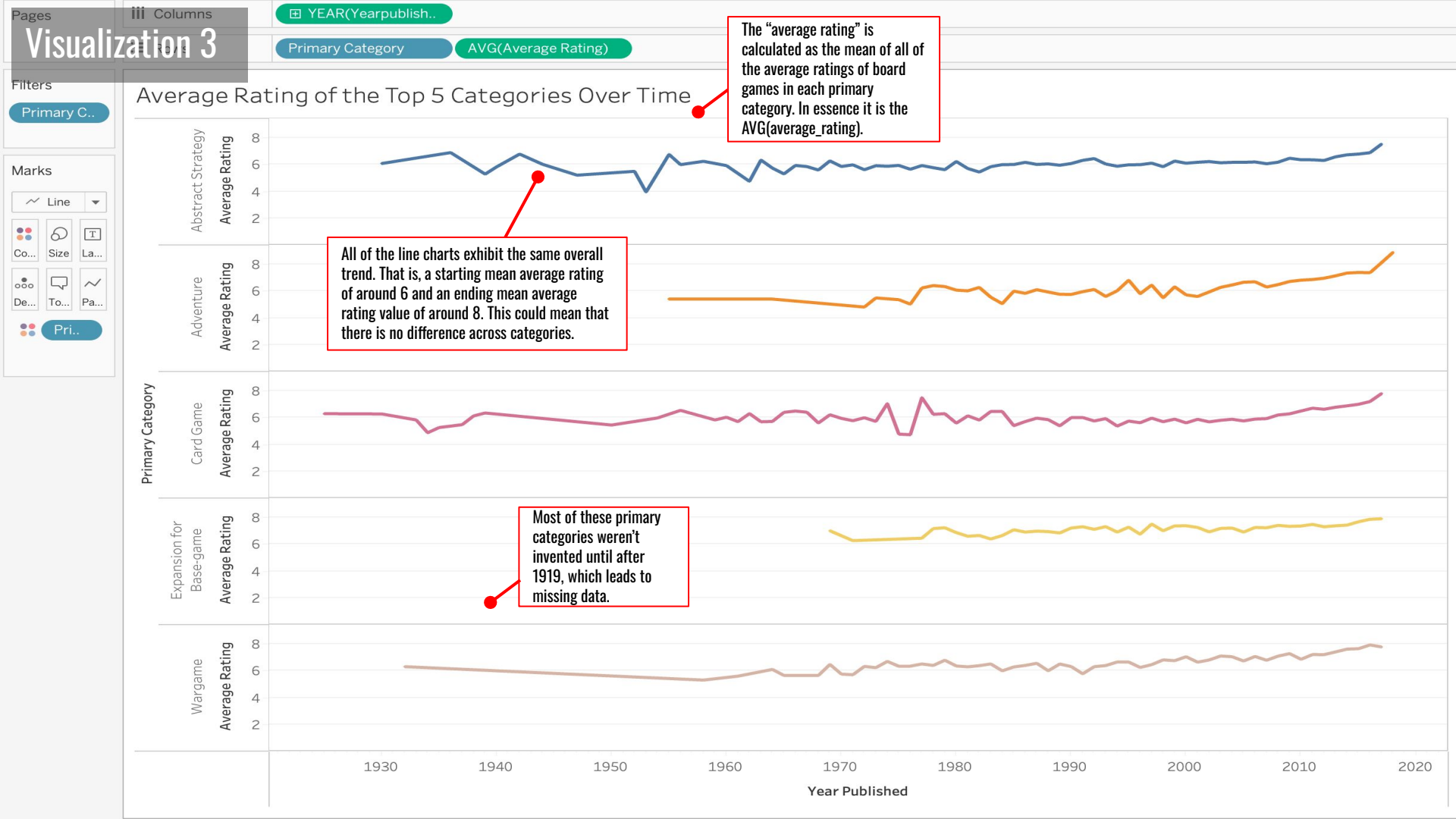
- **Caption:**
  - The pair of boxplots shown above display the different average complexities (average_complexity) on the y-axis for each board game type (game_type) which are displayed on the x-axis. The two types of board games are the original board game (boardgame) and expansions for board games (boardgameexpansion). Each boxplot shows the first, second, and third quartiles of the average complexity attribute, as well as any possible outliers (points outside the upper and lower fence).

- **Description:**
  - The goal of this visualization is to understand the distribution of the average complexity of board games (average_complexity) across the different board game types (game_type). In particular, this visualization answers question 1 from the above investigative questions list. As can be seen from the above boxplots, the board game type called “boardgame” has a slight right skew in its distribution but can be considered approximately normally distributed with its center/median around an average complexity value of 2. Furthermore, the boxplot for the original board game has several high outliers. Also, the board game type called “boardgameexpansion” has a slight right skew in its distribution but can be considered approximately normally distributed with its center/median around an average complexity value of 2.4. Similarly to the other boxplot, the boxplot for board game expansions also has a few high outliers, as well as a single low outlier. Both boxplots have a similar spread, however, the boxplot for the board game expansion has more density in its lower tail than that of the original board game type.

- **Insights/Takeaways:**
  - From the above boxplots and accompanying description, we can see that, on average, the average complexity for expansions is slightly higher than that for the original board games. This makes sense intuitively, because expansions are meant to enhance/add more elements to the game, which in turn makes the game more complex and difficult.



Visualization 3

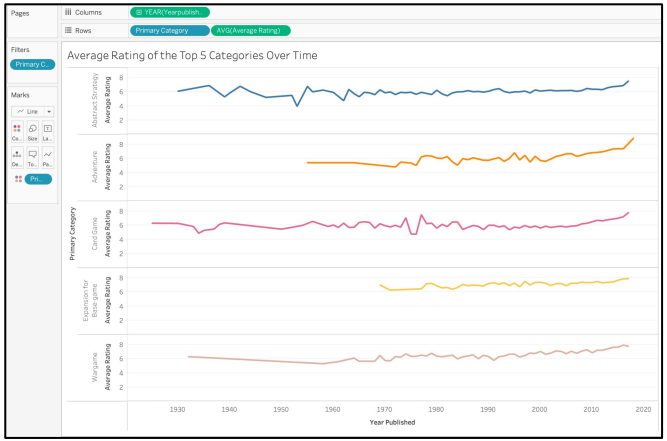


Visualization 3 Analysis

- **Caption:**
  - The five line charts shown above encode the average game rating (average_rating) on the y-axis for each of the top 5 primary game categories (category) in terms of frequency (also shown on the y-axis). The “average rating” on the y-axis is calculated as the mean of all of the average ratings in each of the given game categories. Furthermore, these average ratings per category are expanded along the x-axis due to its encoding of the publishing year (yearpublished).

- **Description:**
  - The goal of this visualization is to understand how the average rating (average_rating) of board games across the top 5 primary game categories (category) changes over the time period of 1919-2019. In particular, this visualization answers questions 1 and 3 from the above investigative questions list. Furthermore, the above line charts show how the average rating of the games released in each year since 1919 have changed over the past century for each of the top 5 most recurring game categories. These top 5 categories are: Abstract Strategy, Adventure, Card Game, Expansion for Base Game, and Wargames.

- **Insights/Takeaways:**
  - Since there are 85 unique primary game categories in this dataset, that would be far too many categories to visualize in a line chart. Thus I decided to only focus on the top 5 categories based on frequency. As can be seen from the above line charts and the accompanying description, each of the top 5 categories of board games has seen an increase of average rating from their first inception, each starting with an average rating of around 6, and peaking/ending with a maximum average rating of around 8 . One other important thing to notice about this graph is that the different categories start at different points in time. For example, the Expansion for Base-game category didn’t have its first game published until the late 1960s, while the Card Game category has been around since the beginning of our allotted time frame. One reason which might explain why the other categories performed better than the Card Game category over all of the years, is that card games have been around for centuries, thus the creation of new categories of games was more exciting to game enjoyers and thus led to higher average ratings for the newer categories, and a stagnating average rating for the Card Game category. The main insight to be gained from this visualization is that there doesn’t appear to be much difference between the average rating over time amongst the top 5 categories. The change appears to be similar across all off the top 5 categories.



Visualization 4

Pages

Filters

Decade

Marks

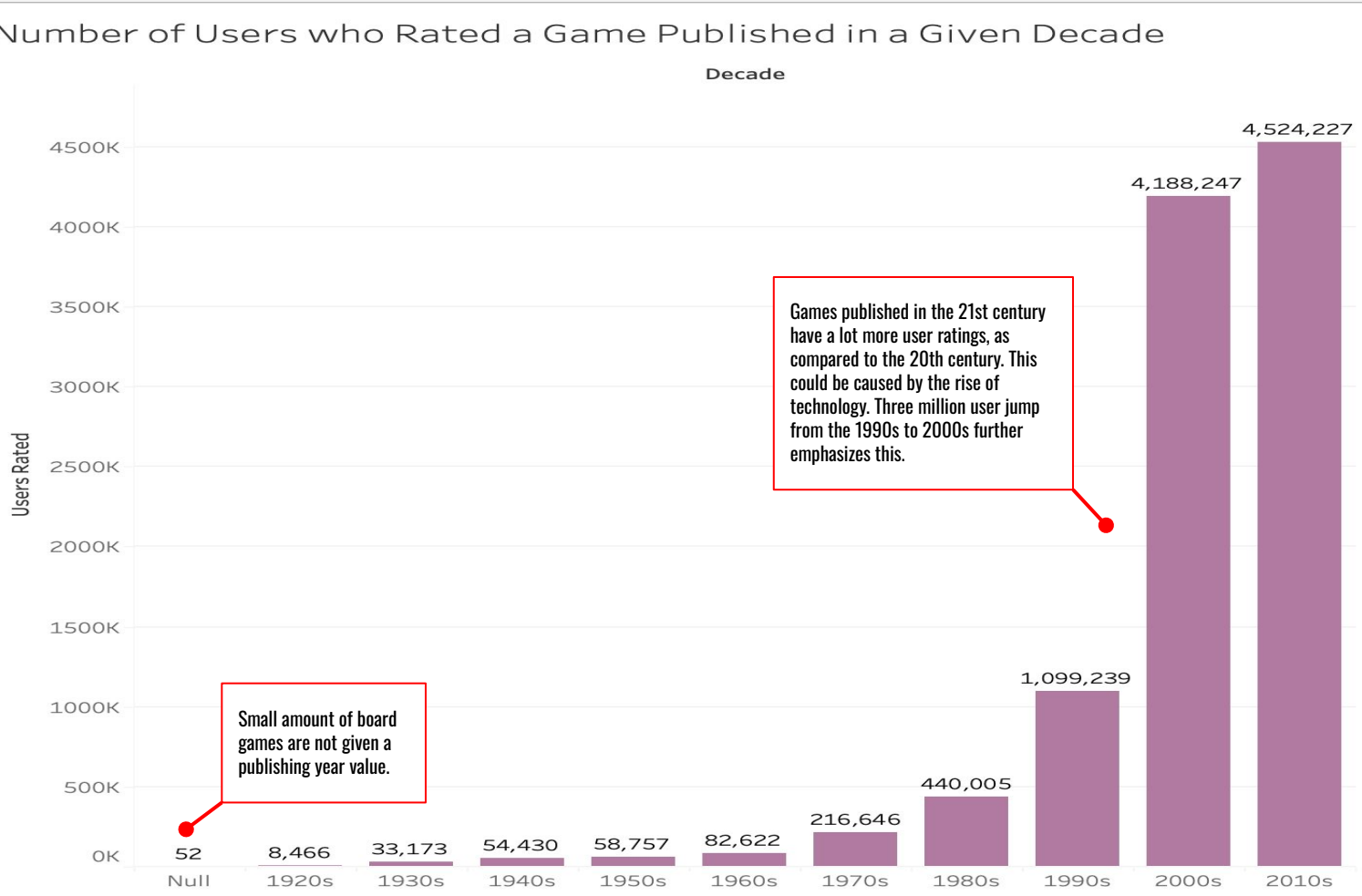
Automatic

Color | Size | Label | Detail | Tooltip

SUM(Users Ra..

Columns: Decade

Rows: SUM(Users Rated)

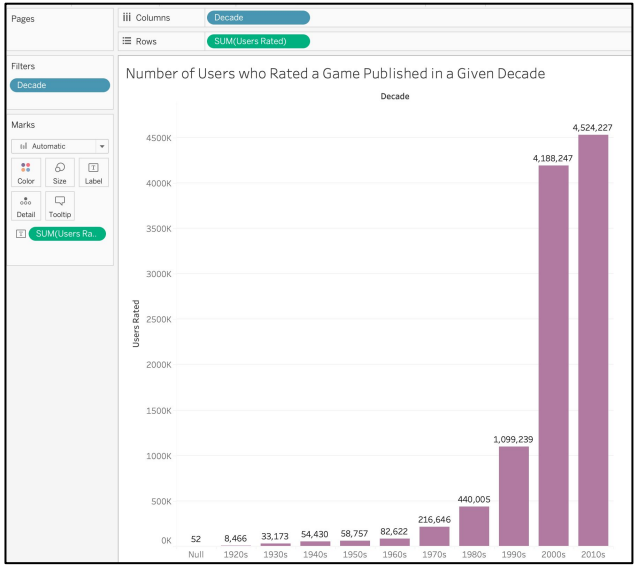


Visualization 4 Analysis

- **Caption:**
  - In the above barplot, the number of users who rated games published in each decade is shown. The total number of user ratings is encoded on the y-axis by the height of the bar, which is the sum of the users_rated variable throughout all of the games published in a given decade. On the x-axis, the decades from the 1920s to the 2010s are shown. These decades were found by recoding the yearpublished variable.

- **Description:**
  - The goal of this visualization is to understand how the number of users rating games (users_rated) relates to the decade in which the game was published (yearpublished). In particular, this visualization answers question 2 from the above investigative questions list. As can be seen from the above barplots, as the decades increase, the number of users rating games published in that decade increases. There also seems to be a few user ratings for games that are given no publishing year designation.

- **Insights/Takeaways:**
  - As can be seen from the above barplot, the number of users rating games from each decade has increased dramatically for each decade since the 1960s. The most dramatic rise was from the 1990s to the 2000s, where an astounding 3 million more ratings were received from games released in the 2000s than were received for games released in the 1990s. Despite such a jump from the 1990s to 2000s, the rise from the 2000s to the 2010s wasn't nearly as big, with only a rise of around 250,000 more ratings. The reason for such a disparity between the amount of ratings for games published in the 21st century versus game published in the 20th century boils down to one simple cause: the rise of technology in the early 21st century. Due to the fact that BoardGameGeek.com didn't even get founded until 2000, mean that most of the major fans of the games from the 1950s-1980s were not living in a technological age in which they could rate games online. Thus, when the site was created, they were past the point where they would feel inclined to rate a game. On the contrary, for every game released in the 21st century, there was an easily accessible place to rate games, and hence why we see an astounding increase in the number of ratings for games released in these decades.



Visualization 5

Pages

Columns: YEAR(Yearpublish..

Rows: Maximum Player Ran..

Filters

Yearpublished

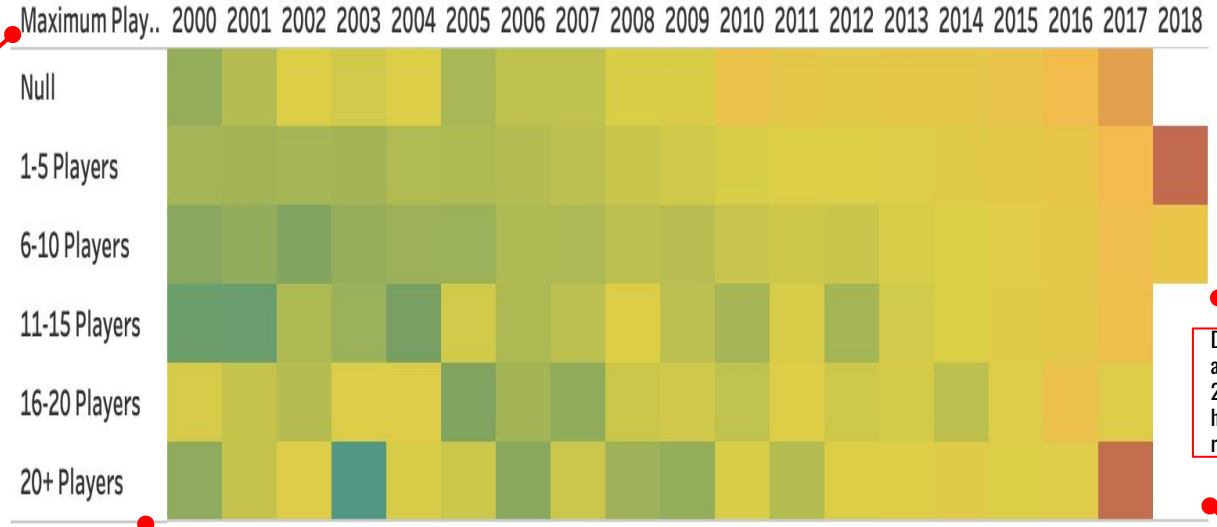
Marks

Automatic

Color | Size | Label | Detail | Tooltip

AVG(Average ..

Heat Map of the Average Rating By Maximum Player Range (2000-2018)



Average of Average Rating

4.574 — 8.917

Maximum Player Ranges

The "average rating" is calculated as the mean of all of the average ratings of board games in each player range. In essence it is the AVG(average_rating). Although the range of the variable is 0.0-10.0, this range only considers the min and max value of this subset of the data.

Due to this lack of data, the mean average rating values in the year 2018 (and even 2016-2017) are highly skewed to the few user ratings available.

This column represents the maximum player ranges of a given board game. That is, the maximum number of players allowed to play the given game. This ensures that a numerical variable is then categorical in order to make this map make sense.

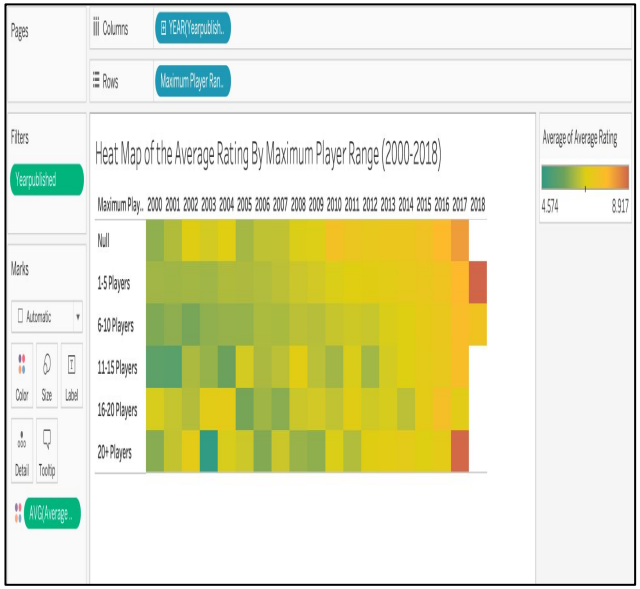
The lack of data for 2018 is due to our user rating > 25 constraint. Newer games have less of a chance to have enough ratings to be in our dataset.

Visualization 5 Analysis

- **Caption:**
  - In the heatmap shown above, the y-axis encodes the ranges of maximum players (maxplayers) of all board games published in the 21st century. As mentioned in the previous sentence, the x-axis encodes the publishing years of board games (yearpublished) for the years 2000-2018. Furthermore, in each of the resulting squares of the grid, the average rating (average_rating) of board games published in the given year with given amount of maximum players allowed is given as a color. These color values are based upon a scale that goes from green to red, which corresponds to values that range from low to high. That is, green represents lower values of average rating, and red represents higher values of average rating. The “average rating” color coded in the grid is calculated as the mean of all of the average ratings for each of the given maximum player ranges for each given publishing year.

- **Description:**
  - The goal of this visualization is to understand how the average rating (average_rating) of games changes throughout the 21st century for different maximum player ranges (maxplayers). In particular, this visualization answers question 2 from the above investigative questions list. As can be seen from the above heatmap, for all maximum player ranges, the average ratings of games increases as the years go on. This result also holds for board games that weren't given a maximum players value in the original dataset. Although this result shouldn’t be surprising as will be explained in the next section.

- **Insights/Takeaways:**
  - Since there are 100 unique publishing years in this dataset, that would be far too many years to visualize in a heatmap. Thus, I decided to only focus on more recent years; those years being the 21st century. As can be seen from the above heat map and the accompanying description, there doesn’t appear to be much difference between the average ratings across games with differing maximum player ranges. Instead, the average game rating seems to be somewhat similar for all maximum player ranges. This means that the maximum player value of a game doesn’t influence the average rating of a game that much. This isn’t too surprising as it is the contents of the game itself that make players excited to play a game/rate it highly, not the number of players that can play the game.



Filters

YEAR(Yearpublished)

Visualization 6

Marks

Pie

Color | Size | Label

Detail | Tooltip | Angle

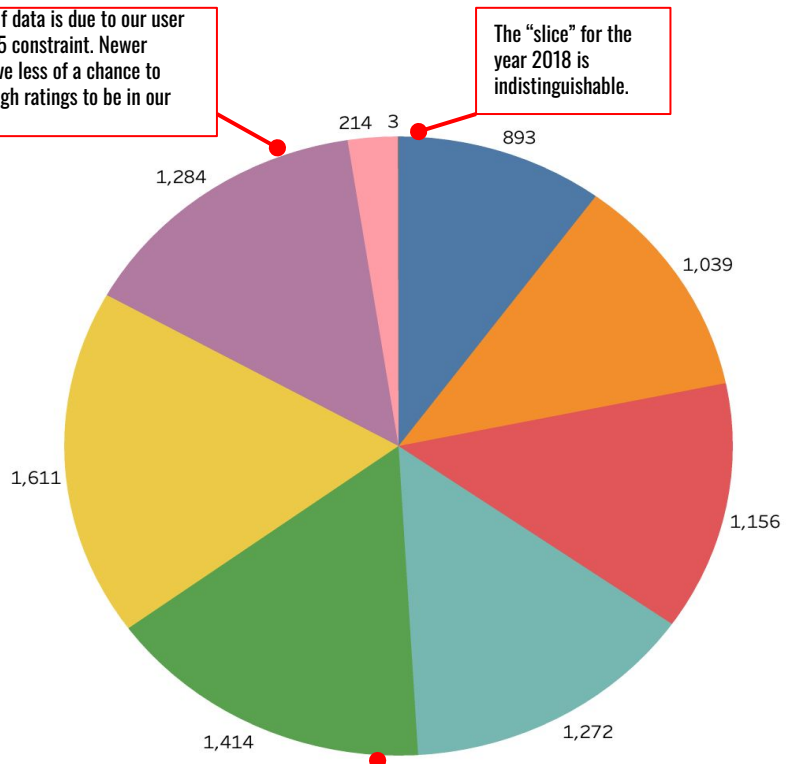
YEAR(Yearp..

CNT(Game Id)

CNT(Game Id)

CNT(Game Id)

Pie Chart For Total Amount of Games Published in Each Year (2010-2018)



Year Published

- 2010
- 2011
- 2012
- 2013
- 2014
- 2015
- 2016
- 2017
- 2018

Total Count

8,886

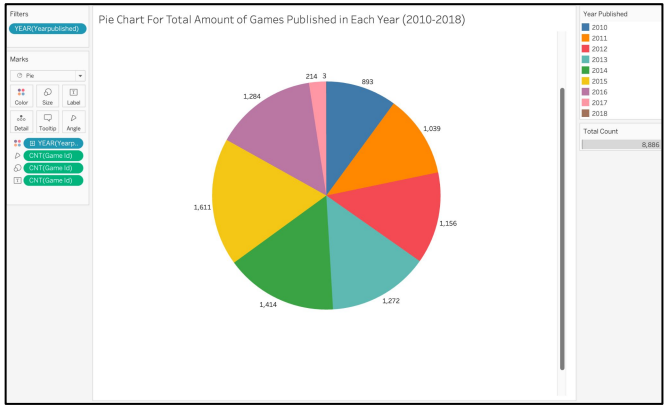
Total count is given for reference.

Visualization 6 Analysis

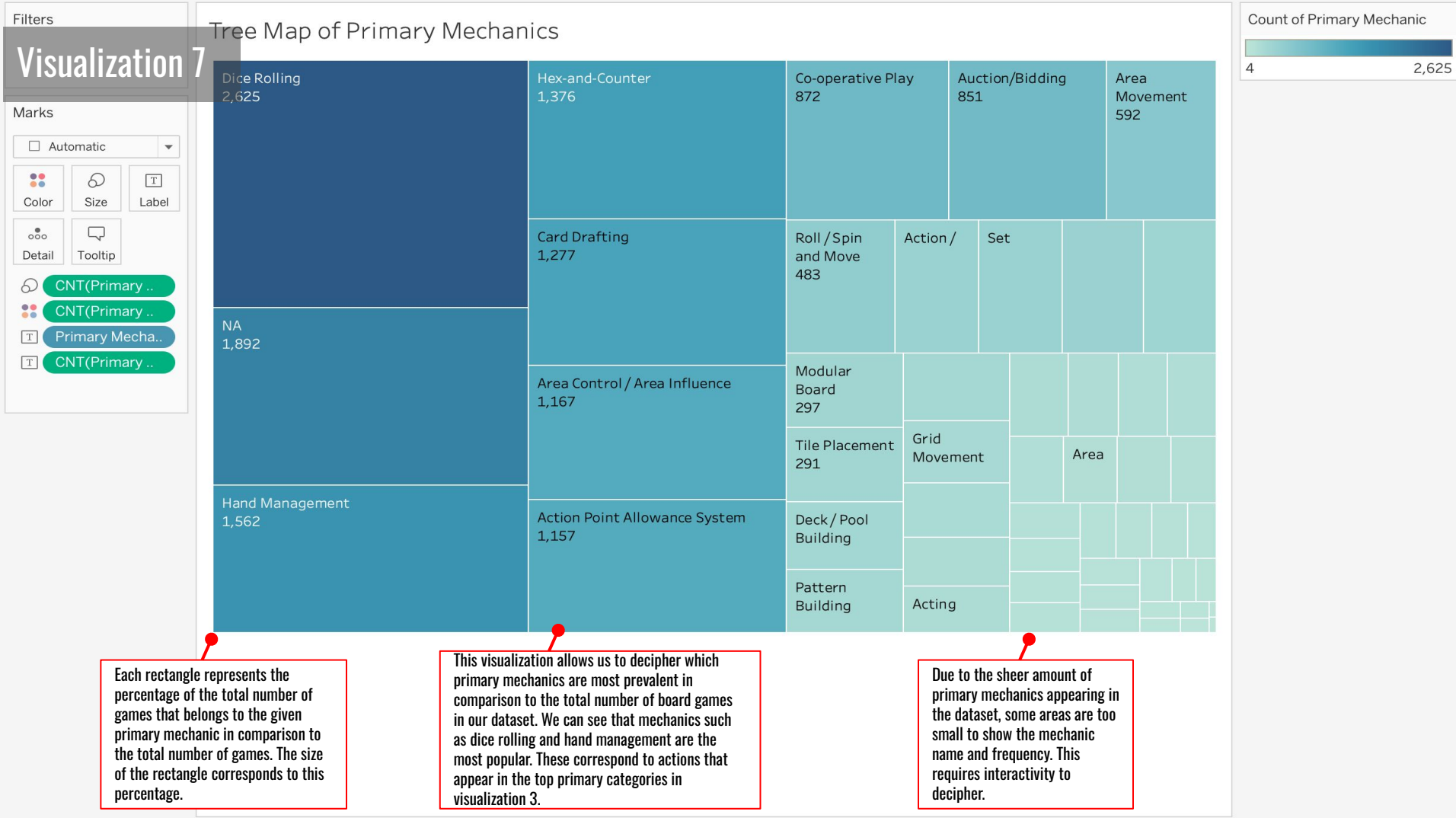
- **Caption:**
  - In this pie chart, the area of the circle represents the total amount of games published in the years 2010-2018. Furthermore, each slice of the pie chart corresponds to a different publishing year (yearpublished), and its area corresponds to the percentage of the total amount of games published that the specific category contributes to. Furthermore, in each category/area, the exact number of games published in that year is also displayed. Each publishing year is given its own unique color, which is determined and displayed in the legend beside the title.

- **Description:**
  - The goal of this visualization is to understand how the number of games published in each year changes throughout the 2010s. In particular, this visualization answers question 2 from the above investigative questions list. As can be seen from the above pie chart, as the years increase, the number of games published in that year also increase. However, for the years 2016-2018, this trend is reversed, which may be due to a lack of recent data. One issue with this pie chart is that the value associated with the year 2018 is so small that it cannot be visualized as a slice in the pie chart.

- **Insights/Takeaways:**
  - Since there are 100 unique publishing years in this dataset, that would be far too many years to visualize in a pie chart. Thus I decided to only focus on more recent years; those years being the 2010s. As can be seen from the above pie chart and the accompanying description, there seems to be more games published as the years go on. This phenomenon makes sense due to the fact that innovations in technology expand the definition of board games and allow for new ideas to come to fruition. Another important takeaway is that more recent games have less reviews on them. This result is apparent when looking at the years 2016-2018. In particular, the trend that we saw throughout the entirety of the pie chart reveres, this can be explained by the fact that we filtered the data for only the board games that had 25 or more ratings. Thus, newer games that didn't have time to gain much exposure were omitted.



Visualization 7



Each rectangle represents the percentage of the total number of games that belongs to the given primary mechanic in comparison to the total number of games. The size of the rectangle corresponds to this percentage.

This visualization allows us to decipher which primary mechanics are most prevalent in comparison to the total number of board games in our dataset. We can see that mechanics such as dice rolling and hand management are the most popular. These correspond to actions that appear in the top primary categories in visualization 3.

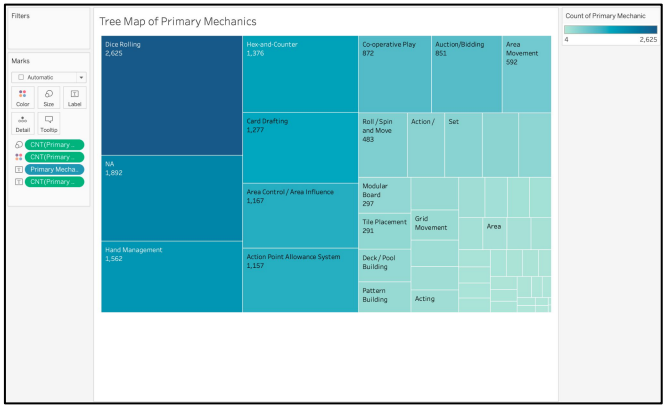
Due to the sheer amount of primary mechanics appearing in the dataset, some areas are too small to show the mechanic name and frequency. This requires interactivity to decipher.

Visualization 7 Analysis

- **Caption:**
  - In this tree map, as Tableau calls it, there is no notion of axes. Instead the notion of areas is instead adopted. In particular, the area/entirety of the rectangle as a whole represents the total amount of games published between 1919 and 2018 (game_id). Furthermore, the big rectangle is further broken up into smaller rectangles, which represent the 51 different mechanics (mechanic) some of which are too small to be visualized. The size of each of these rectangles corresponds to the percentage of the total amount of games published that the mechanic corresponds to. In each smaller rectangle, the name of the mechanic and the amount of games published with that primary mechanic are displayed (if the size of the rectangle allows it). Lastly, the color of the rectangle also corresponds to the the amount of games published with that primary mechanic.

- **Description:**
  - The goal of this visualization is to visualize which primary mechanics (mechanic) appears most frequently in board games, and how these frequencies compare to the grand total. In particular, this visualization answers question 3 from the above investigative questions list. As can be seen from the above tree map, mechanics such as: Dice Rolling, Hand Management, Hex-and-Counter, Card Drafting, etc. It is also apparent that quite a few of the games in the dataset are not given a primary mechanic and are thus put as NA. These most frequent mechanics will be further analyzed in the next visualization. Lastly, this visualization is majorly flawed as primary mechanics that are less frequent do not have enough space to display its name and frequency, thus users must rely on interactivity, which isn't an option in static visualizations.

- **Insights/Takeaways:**
  - As can be seen from the aforementioned tree map and accompanying description, we can see that primary mechanics such as Dice Rolling, Hand Management, Hex-and-Counter, Card Drafting are the most frequent. These mechanics are fundamental to most games that are created so it is not surprising that they are the most frequent. One important connection to the primary categories described in visualization 3, is that many of these mechanics are present in the most frequent categories. For example, card games are the most frequently created game category, and these games include the mechanics Hand Management and Card Drafting. Therefore it appears that there is a connection between the top primary game categories, and the top primary game mechanics.



Visualization 8

Pages

Columns: Primary Mechanic | Average Complexity

Rows: Average Rating

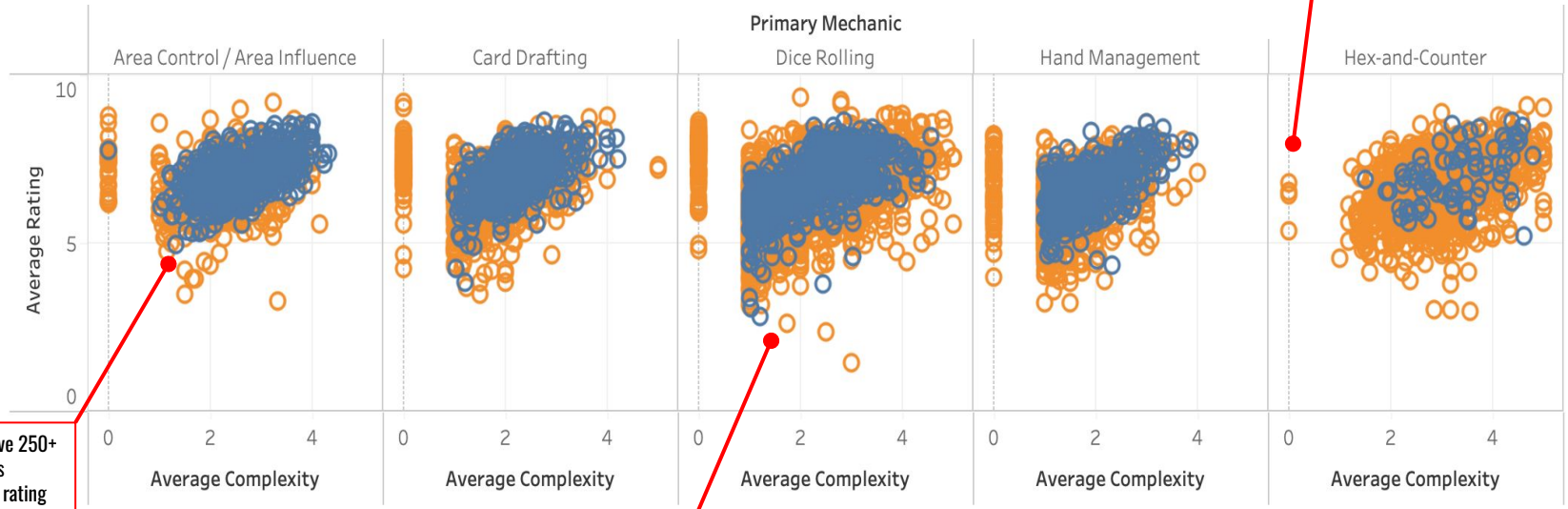
Filters: Primary Mechanic

Users Rated Range: 250+ <250

Legend might be hard to see due to its unusual location.

Marks: Automatic; Color, Size, Label, Detail, Tooltip, Shape; Users Rated Ra..

Average Rating Versus Average Complexity For the Top 5 Mechanics



Quite a few board games are not given an average complexity score. These are represented by the vertical bars at x=0 in the plots.

Board games that have 250+ user ratings have less variability in average rating across all values of average complexity. This implies that highly reviewed games are more "reliable."

All of the scatter plots exhibit the same positive linear relationship between average rating and average complexity. This could mean that there is no difference in the relationship across mechanics.

Visualization 8 Analysis

- **Caption:**
  - The five scatterplots shown above encode the average rating (average_rating) value on the y-axis and the average complexity (average_complexity) value on the x-axis for all board games published since 1919. These data points are plotted as dots on the Cartesian plane. Furthermore, the horizontal axis also encodes the top 5 mechanics. In particular, there is a scatterplot made for each of the top 5 mechanics. Furthermore, all of the data points in all of the top 5 mechanics are broken up into two types: those that have over 250 ratings, colored in blue, and those that have under 250 ratings, colored in orange.

- **Description:**
  - The goal of this visualization is to show the relationship between the average rating (average rating) and the average complexity (average_complexity) of all board games published since 1919 for each of the top 5 game mechanics. These mechanics include: Area Control/Area Influence, Card Drafting, Dice Rolling, Hand Management, and Hex-and-Counter. In particular, this visualization answers questions 3 from the above investigative questions list. As can be seen from the above scatterplots, all of the individual mechanics have the same moderate positive linear association between the average game rating and average complexity variables. The two game mechanics that have the strongest positive linear association are the Card Drafting and Hand Management mechanics. While the Dice Rolling mechanic had the weakest association between these two variables. Furthermore, on average, the points that had the more ratings usually were higher rated than there lesser rated counterparts. Lastly, games that have a high number of user reviews tend to have a larger variance in their average_rating value across all values of average_complexity.

- **Insights/Takeaways:**
  - Since there are 51 unique primary game mechanics in this dataset, that would be far too many mechanics to visualize in a scatterplot. Thus, I decided to only focus on the top 5 mechanics based on frequency. Thus, based on the above scatterplots and accompanying description, we can see that, in general, as the average complexity of a game increases, the average rating of the game also increases, independent of the primary game mechanic of the game. This can mainly be explained due to the fact that board games themselves are a niche hobby to begin with. Since board games aren't the most popular pastime among other things like video games, movies, TV shows, etc. they have become more of a pastime for a particular and small audience. Thus, those specialized audiences tend to enjoy more complex games in order to challenge themselves. Hence, on average, when a game gets more difficult, avid board game fans accept the challenge and thus have higher ratings for the game. Also, based on the fact that the games with a lot of user reviews have lower variability in average_rating scores across all values of average_complexity, it follows that games with more user reviews tend to have more accurate and reliable information about the given game as opposed to those with a smaller amount of user ratings.

Visualization 9

Pages

Columns: YEAR(Yearpublish..

Rows: Primary Category

Filters

- Yearpublished
- Primary Category

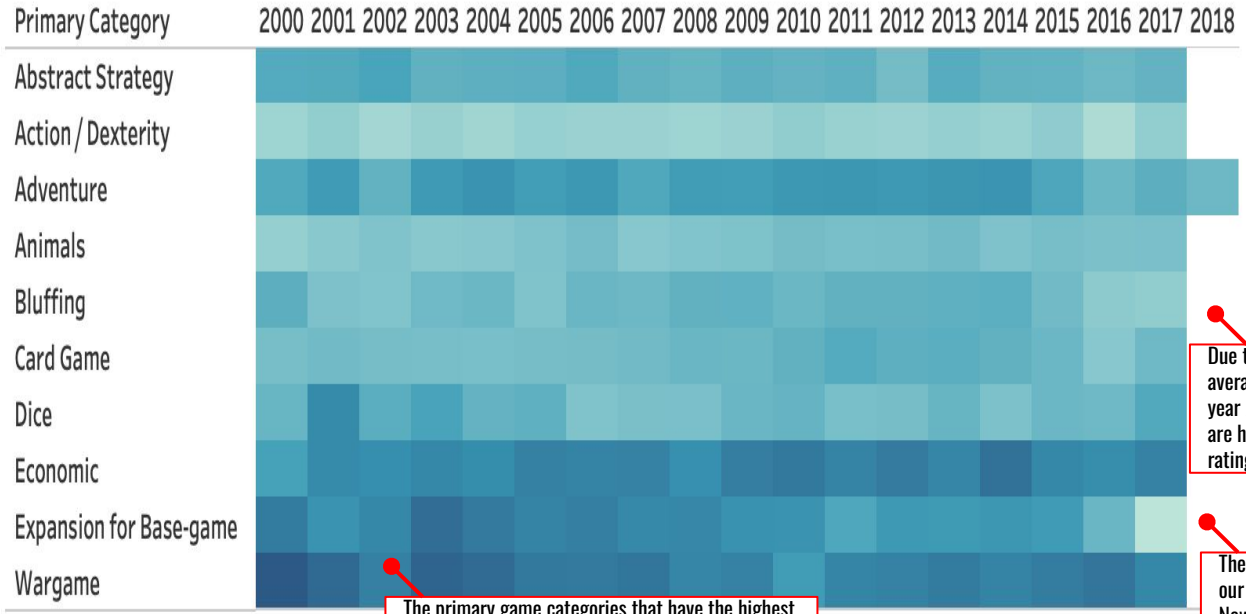
Marks

Automatic

Color | Size | Label | Detail | Tooltip

AVG(Average ..

Heat Map of the Average Complexity For the Top 10 Categories (2000-2018)



Average of Average Complexity

0.813 — 3.588

The "average complexity" is calculated as the mean of all of the average complexities of board games in each player range. In essence it is the AVG(average_complexity. Although the range of the variable is 0.0-5.0, this range only considers the min and max value of this subset of the data.

Due to this lack of data, the mean average complexity values in the year 2018 (and even 2016-2017) are highly skewed to the few user ratings available.

The lack of data for 2018 is due to our user rating > 25 constraint. Newer games have less of a chance to have enough ratings to be in our dataset.

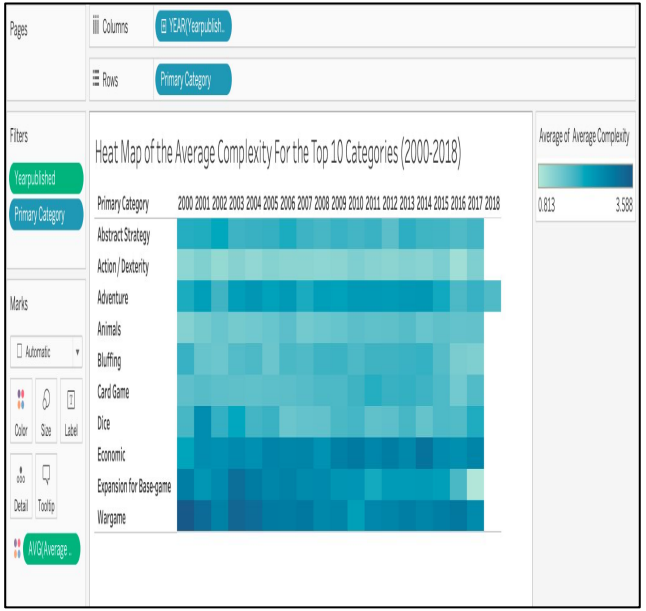
The primary game categories that have the highest mean average complexity value are expansions, economic, and wargames. This result is expected since wargames and strategy games are often seen as complex by the general public..

Visualization 9 Analysis

- **Caption:**
  - In the heatmap shown above, the y-axis encodes the top 10 primary game categories (category) of all board games published in the 21st century. As mentioned in the previous sentence, the x-axis encodes the publishing years of board games (yearpublished) for the years 2000-2018. Furthermore, in each of the resulting squares of the grid, the average complexity (average_complexity) of board games published in the given year in each given category is given as a color. These color values are based upon a scale that goes from light blue to dark blue, which corresponds to values that range from low to high. That is, light blue represents lower values of average complexity, and dark blue represents higher values of average complexity. The “average complexity” color coded in the grid is calculated as the mean of all of the average complexities for each of the given primary categories for each given publishing year.

- **Description:**
  - The goal of this visualization is to understand how the average complexity (average_complexity) of games changes throughout the 21st century for the most popular game categories (category). In particular, this visualization answers questions 2 and 3 from the above investigative questions list. As can be seen from the above heatmap, primary categories such as: Wargame, Economic, Abstract Strategy, Adventure, and Expansion for Base-game all have higher mean average complexity scores across the 21st century as compared to primary categories such as Action, Animals, and Card Games. It is also important to note that these mean average complexity values remain relatively constant for each primary category.

- **Insights/Takeaways:**
  - Since there are 100 unique publishing years in this dataset, that would be far too many years to visualize in a heatmap. Thus, I decided to only focus on more recent years; those years being the 21st century. Furthermore, since there are 85 unique primary game categories in this dataset, that would be far too many categories to visualize in a heat map. Thus I decided to only focus on the top 10 categories based on frequency. As can be seen from the above heat map and the accompanying description, there does appear to be differences in the mean average complexity across the top 10 most frequently published game categories. This is expected, as it is general consensus that strategy and wargames are more complex than dice and card games (for the most part). One result that is surprising is that average complexity seems to be independent of publishing year. I would’ve expected that games would get more complex over time.



Assignment Reflection

How did you approach the design process?
I approached the design process of making my visualizations in a unique way. In particular, unlike my visualizations in assignment 2, I tried to make a unique visualization type for all 8 of the required visualizations. Furthermore, I wanted to make use of a wide variety of the attributes that the dataset provided in order to learn more about board games themselves. With these two goals in mind I made unique and diverse visualizations that attempted to answer the questions mentioned in the next section.

What question(s) did you attempt to answer with your visualizations?
As mentioned in part 1 of the assignment, the questions I attempted to answer with these visualizations are:

- How do the two types of board games compare? That is, what are the similarities and differences between board games, and board game expansions?"
- How have certain aspects and features of board games changed over time? Has there been any noticeable changes, or have these aspects and features remained relatively constant?
- How do certain aspects and features of board games compare between the top categories and mechanics of board games?

Visualizations 1 and 2 attempted to answer question 1, visualization 3-6, and 9 attempted to answer question 2, while visualizations 3 and 7-9 attempted to answer question 3.

What did you struggle with?
One thing that I struggled with during this assignment, as described in part 1, was that the full dataset was running very slow in Tableau. In order to combat this I was forced to trim the dataset by almost 70,000 rows. In the end, this data trimming allowed for more clear analysis, so the situation wasn't all bad. Another aspect of this assignment that I struggled with was using Tableau itself. In particular, I struggled with some of the more complicated aspects of Tableau that were not explicitly mentioned in class.

What did you enjoy?
The main thing that I enjoyed about this assignment is that I was able to work with a dataset that I got to pick (although the selection of this dataset was one of the most time consuming and challenging parts of the whole assignment). Furthermore, the ability to choose which dataset I worked with allowed me to learn information about a topic that is close to my heart. In particular, it was fulfilling to learn this information through visualizations that I thought of and curated. Lastly, due to the fact that this data has a bit of sentimental value, I was more motivated to explore the data than I would have been if I had to use a pre-selected dataset.

What did you learn?
The main takeaway that I gained from this assignment was using the Tableau GUI. Since I had never used a graphical user interface like Tableau before, it was difficult for me to get a hang of how I was supposed to make the specific visualizations that I wanted, especially compared to programming languages such as R and Python. By the time I reached the final few visualizations, I has a good idea of what I needed to do in order to get the visualization to appear exactly how I wanted them to. One of my proudest achievements in this assignment was my creative use of calculated fields in order to make my visualizations display more information than they would've been able to otherwise.