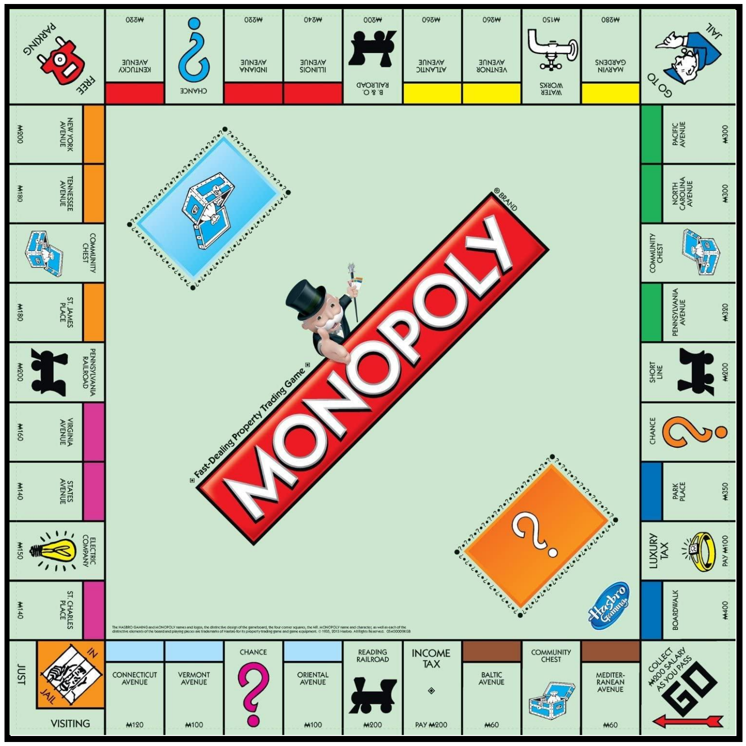
I. INTRODUCTION

Saying that Monopoly was a game I played a lot growing up, is an understatement. I have always been enamored by every aspect of the game: logistics, real estate, mathematics, competitiveness, etc. This drove me to always seek ways in which I could increase my chances of winning when I played the game with family and friends. I would take on the role of being the banker so that I can keep track of everyone's in-game finances, or save my money so that I could purchase the most expensive properties with high rents. However, I noticed that despite my attempts to strategize, I was not winning any more often. I've always wanted to know the reason why, and so I am conducting this data visualization analysis.

Before beginning the analysis, here is a brief background on the rules and premise of Monopoly: It is a 2-8 player game where each player is given a fixed amount of money at the start ($1500). Then, they take turns rolling the dice and moving clockwise around the board as dictated by the sum of the numbers on the pair of dice. There are three types of spaces on a Monopoly board:

> Property: If a player lands on a property, they can decide whether or not they would like to purchase it. After the purchase, every other player who lands on the square needs to pay a certain amount of rent as dictated by the property deed.
> Chance and Community Chest: There are three Chance and Community Chest spaces each on the entire board. Both require the player to pick a card from the specified deck and perform the action listed on it. While Community Chest cards only change the finances of the player, Chance cards have the ability to transport the player to another space on the board within the same turn.
> Corners: Each corner on the Monopoly board has a distinct feature. The first corner is "Go" which marks the beginning and end of one traversal, and the player collects $200 each time it is passed. The second corner is "Jail", where players can be sent in many ways. The third corner is "Free Parking" which has no action, while final corner is "Go to Jail", where every player landing on it is directly sent to Jail.



The objective of the game is to manage one’s finances well and eventually bankrupt all the other players of their money and properties, effectively creating a real estate monopoly on the board.

II. DATASET

Collecting data from a game of Monopoly is unlike formal data collection processes. Since the outcomes within the game are governed by the game design as well as random processes like dice rolls and card draws, the nature of any data that can be collected is simulated. For this project, I used the publicly available *Monopoly Board Frequencies and Economies* dataset from data.world, created by contributor Ben James (Dataset Link).

The dataset is based on the premise that a single-player performs all the traversal operations within the game without purchasing any property, for exactly 60000 turns without the interference of any other player. This data is collected by running a computer simulation of this premise. While this means that the interactions between the actions of different players cannot be studied, the long-term trends of the underlying game mechanics can be studied through data visualization. The large number of observations in this dataset allows us to make the assumption that each of the turns is independent events - In reality, each turn is dependent on all of its predecessors. However, the expectation is that with a large dataset all the paths traversed are explored frequently enough to provide an accurate representation of trends when the game of Monopoly is played to completion.

	Turn	Card Draw	Die Roll 1	Die Roll 2	Sum of Dice	Board Position	Doubles	3 Straight Doubles	Space Name
0	1	NaN	1	5	6	0	NaN	NaN	Go
1	2	NaN	2	6	8	6	NaN	NaN	Oriental Ave
2	3	NaN	3	1	4	14	NaN	NaN	Virginia Ave
3	4	NaN	4	2	6	18	NaN	NaN	Tennessee Ave
4	5	NaN	6	4	10	24	NaN	NaN	Illinois Ave

Throughout this analysis, we will be using subsets of this main dataset as well as accompanying data dictionaries, like the Chance and Community Chest Card Information, Space Names and their indices as well as Property rents.

III. METHODOLOGY AND INSIGHTS

(a) SPACE FREQUENCY HEATMAP

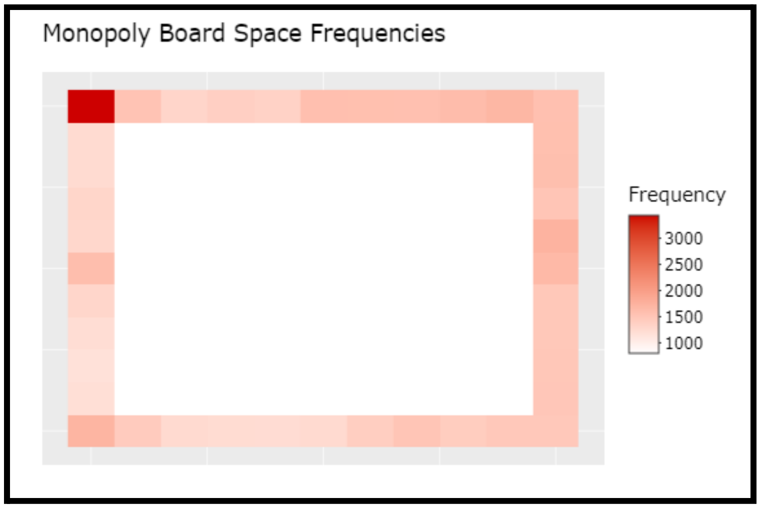
With a large dataset recording a sequence of turns and their starting position, the first question to arise is which spaces are more frequented than others. In order to answer this question, we require the entire dataset to be grouped by Space Name while maintaining the counts of observations within each group:

	Space Name	Frequency	Index
0	Atlantic Ave	1491	26
1	B & O Railroad	1671	25
2	Baltic Ave	1225	3
3	Boardwalk	1443	39
4	Chance #1	1316	7

Once this is obtained, it is necessary to decide the type of visualization that is appropriate for this data. We need to represent the frequency (numeric variable) of each board space (categorical variable). With 40 unique categories, a bar chart is not a feasible option to represent all of the data. Also, the data in the Board Space categorical variable are also spatially related to one another. The chosen chart type must retain the relationship between the categories while encoding the frequency.

A heatmap is the best choice. With the 2D grid structure available in a heatmap, the board spaces can be displayed with integrity to the game. However, since heatmaps encode a square matrix, we need additional data points - The board spaces only include the borders but the internal part of the heatmap also needs to be filled with blank entries to create the visualization.

All of the observations were given x and y coordinates according to their positions on the board. The size of the heatmap is 11 x 11. The internal 9 x 9 square consists of spaces given a frequency of 0, thereby creating the appearance of an actual board shape. The frequency is encoded in the intensity of the red color of the square.



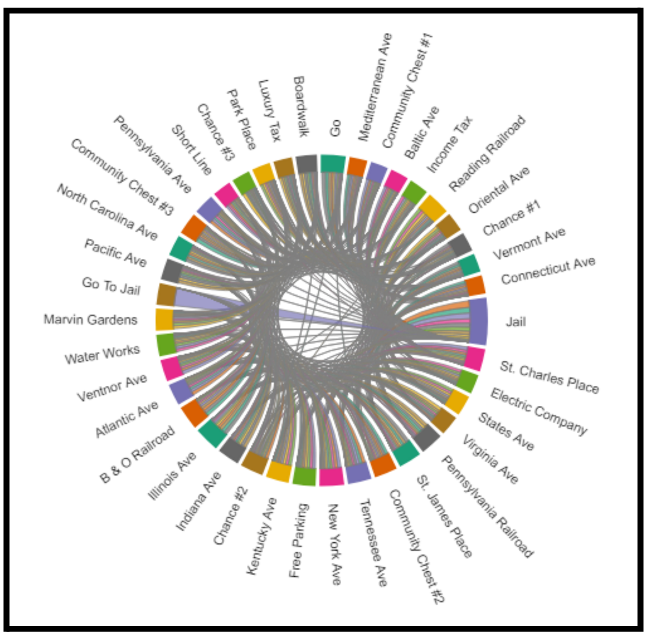
The visualization is also interactive, with an interactive scroll that reveals the following information: Space Name, Frequency. The visualization tasks associated with this visualization are to locate the spaces with high/low frequency and compare the frequency between different observations. This design achieves both of those outcomes as it abstracts all of the numeric data and presents a clean version of the board while encoding information. The design also takes advantage of the pre-attentive processing of color - with intensity being ordinal by intuition, the users of the visualization can immediately identify that the space on the top left corner (Jail) is by far, the most frequented.

(b) BOARD TRAVERSAL CHORD DIAGRAM

The next question to be answered is what are the most common connections between board spaces. In order to answer this question, we require the starting space and ending space for each turn in the dataset. Currently, each observation in the dataset records only the starting space. However, given that the entries are sequential, it is possible to create a copy of the 'Space Name' column, move it up by one index, to create the 'Next Space' attribute. This, of course, reduces the number of observations to 59999.

	Current	Next	l_next	l_current
0	Go	Oriental Ave	6	0
1	Oriental Ave	Virginia Ave	14	6
2	Virginia Ave	Tennessee Ave	18	14
3	Tennessee Ave	Illinois Ave	24	18
4	Illinois Ave	Pennsylvania Ave	34	24
...	...	...	...	...
59994	Boardwalk	Jail	10	39
59995	Jail	New York Ave	19	10
59996	New York Ave	Water Works	28	19
59997	Water Works	Park Place	37	28
59998	Park Place	Reading Railroad	5	37

Once this is obtained, a multi-index is created on (Current, Next) to find the frequency of any given edge. Now, it is necessary to decide the type of visualization that is appropriate for this data. Here, we have two categorical variables (Current and Next) as well as one numeric variable (Frequency). A chord diagram is a suitable representation for this data. Chord diagrams are generally used to represent network data. Here, each node corresponds to a Board Space, and the connection between two spaces is represented using a connection arc.



The frequency of each connection is encoded as the thickness of the edge. The most common connection exists between Jail and Go to Jail as observed by it being the thickest arc. The nodes are in order of game space index and move clockwise as in the game board, establishing intuition for the user of the visualization. Like in the game board, the Jail and Go to Jail space also diametrically oppose each other in the visualization.

The primary insight in this visualization is that most of the paths between the edges are reached using the dice. Given that the movement through dice rolls is limited from 2 to 12 spaces, most arcs form connections proximal to the source node. Given that the nodes are ordered in the chord diagram, this is represented with the majority of connections being concentrated uniformly around the periphery of the center. The connections spanning a large number of spaces are represented by the sparseness close to the center.

(c) COMMUNITY CHEST COMPARISON BAR CHART

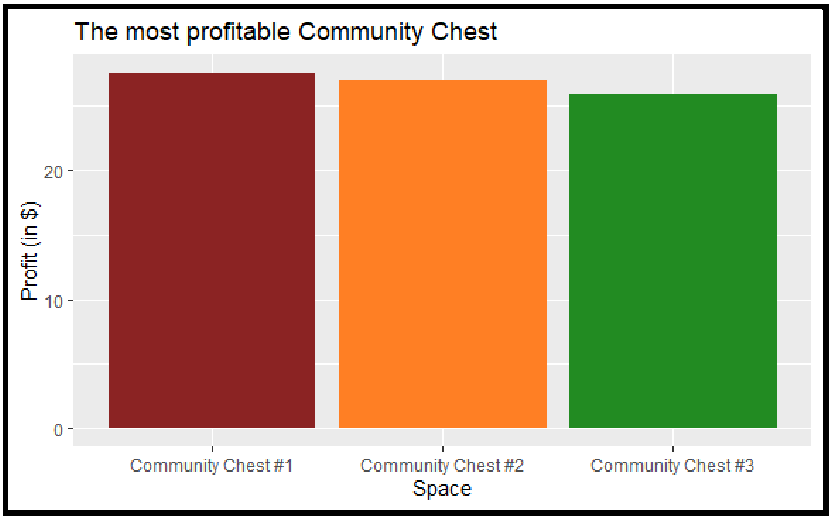
This next section deals with whether the three Community Chest spaces on the board are equally "lucky" or profitable for the player. Unlike Chance cards, these cards never cause a player to move away from their current position and instead affect the finances of the player. Since we have access to all the turns as well as the specific card picked from the deck on those turns, we can calculate profits on these squares. The data pre-processing is as follows:

- Retrieve a subset of the dataset where Space Name contains Community Chest
- Join this dataset with another supplementary dataset that contains information on the monetary change dictated by each of the cards. This left outer join is made on the attribute "Card Draw".

	Turn	Card Draw	Die Roll 1	Die Roll 2	Sum of Dice	Board Position	Doubles	3 Straight Doubles	Space Name	Profit
8	9	6	5	2	7	17	NaN	NaN	Community Chest #2	40
10	11	5	1	6	7	33	NaN	NaN	Community Chest #3	0
21	22	3	3	4	7	2	NaN	NaN	Community Chest #1	-50
38	39	14	1	3	4	17	NaN	NaN	Community Chest #2	10
105	106	16	6	2	8	33	NaN	NaN	Community Chest #3	50

	Space Name	Profit
0	Community Chest #1	27.553282
1	Community Chest #2	26.964624
2	Community Chest #3	25.941409

Now, the last step is to simply group the dataset on "Space Name", which forms three groups (#1, #2, #3), and average the profit within the groups. The hypothesis would have been that since card draws are random processes, the profit on each space should be about the same. From this data, it is evident that the profit on Community Chest #1 is the highest. The range of the Profit variable is about $1.6, which is quite small when considering the currency used in the game. Given that this data has one categorical variable (Space Name) and one numeric variable (Profit), where the different categories are related spatially but are few in number, a bar chart is an appropriate visualization.



The visualization title and axis labels concisely convey information about the visualization task - to observe the differences in average profit of landing on different Community Chest Spaces. Beyond the slight differences in the heights of the bar, the main insight to be conveyed in the visualization is that the values are roughly equal. As a result, the size of a single unit on the y-axis ($10) being too large to closely estimate the height differences is not a point of concern. The bars are also given different colors that correspond to the family of properties within which they are found - different clusters of properties in Monopoly have their own distinct color. While two of the bars are colored shades of red and green, which could raise concerns for the ability of color blind people to observe the contrast, gaining insight from bar charts is not dependent on contrasting any other physical attribute of the bar, except its height. These colors were chosen to convey in a limited and implicit context, where these spaces belong on the board.

(d) MOVEMENT ACTIONS IN CHANCE SANKEY DIAGRAM

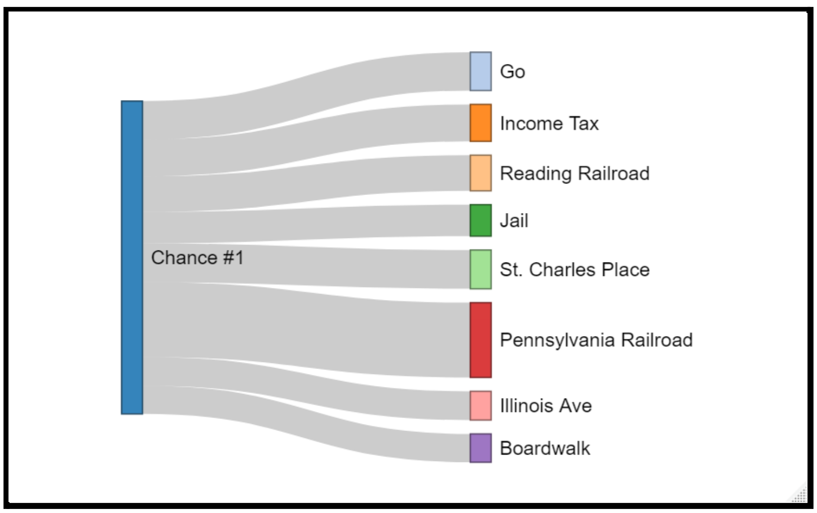
Some actions dictated by drawing Chance cards involve the movement of the player from that space to another on the board. This next section seeks to investigate whether there are some spaces that are visited from Chance, more often than others. In order to analyze flow in more detail, I chose a single Chance board space - Chance #1 and focus on the spaces reachable from it, within the same turn. Here is the data pre-processing:

- Retrieve a subset of the dataset where Space Name contains Chance #1
- Join this dataset with another supplementary dataset that contains information on the movement dictated by each of the cards. This left outer join is made on the attribute “Card Draw”.
- Then, the dataset is grouped on the attribute “Space Name”, and the number of observations in each group is maintained as frequency.

- When making the visualization, the group which records turns that end in Chance #1 (no movement) is removed.

	To	Count	Space Name
0	0	84	Go
1	4	81	Income Tax
2	5	78	Reading Railroad
3	7	630	Chance #1
4	10	69	Jail
5	11	85	St. Charles Place
6	15	164	Pennsylvania Railroad
7	24	63	Illinois Ave
8	39	62	Boardwalk

The spaces under "Space Number" refer to the destination node in that turn. All of the observations have the same source node - Chance #1. The variables of interest to be visualized are: Two categorical variables (Source and Destination) and one numeric variable (Count). As a result, the most appropriate visualization to represent this simple network is a Sankey diagram. A Sankey diagram is typically used to show flow, with the thickness of the edge leaving a node corresponding to greater flow.



The major insight from this visualization is that most of the flows are roughly equal in size - except the edge between Chance and Pennsylvania Railroad. This discrepancy is explained by the fact that there are two Chance cards in the deck that point to Pennsylvania Railroad while only one for every other reachable space.It is extremely interesting that the idea behind creating the game rules is reflected by the simulated data, and more clearly in this visualization.

(e) PROFITABILITY HEATMAP

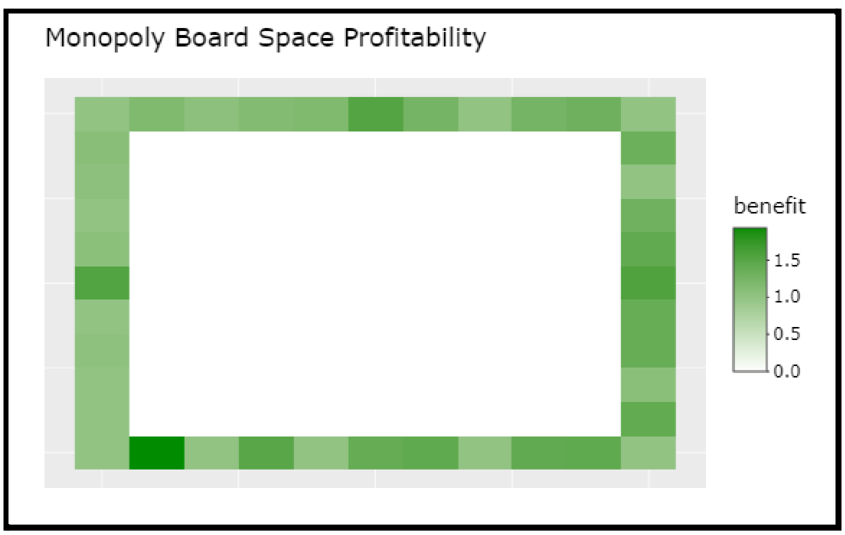
The final question that I will be exploring is how profitable different properties are in comparison to one another. Without a singular defined profit measure in the game, I identified three key characteristics of a property that translates to investment return: Low initial cost, High rent and High probability of landing on the property. Therefore, using the formula:

1 - (Cost of Property / Total Money) + P(Landing on the square) * Base Rent

I calculated the value of this metric for each of the property spaces. Spaces like Jail and Free Parking were disregarded in this analysis, as they cannot be purchased. The only pre-processing involved was to join the simulation dataset to one containing the property cost and rent, using the "Space Name" variable as the key.

	Space Name	Frequency	Index	x	y	Cost	Rent	Proportion Initial Left	benefit
0	Atlantic Ave	1491	26	10	4	260.0	22.0	0.826667	1.373367
1	B & O Railroad	1671	25	10	5	200.0	25.0	0.866667	1.562917
2	Baltic Ave	1225	3	0	3	60.0	4.0	0.960000	1.041667
3	Boardwalk	1443	39	1	0	400.0	50.0	0.733333	1.935833
4	Chance #1	1316	7	0	7	0.0	0.0	1.000000	1.000000

The variables of interest in the processed data are Space Name (categorical) and benefit (numeric). Like with the first visualization question, the most appropriate visualization is a heatmap, created using the same process described above.



The profitability is encoded in the intensity of the green color of the square. The visualization is also interactive, with an interactive scroll that reveals the following information: Space Name, benefit. The visualization tasks associated with this visualization are to locate the spaces with high/low benefit and compare the profitability of different spaces. The user of the visualization is immediately able to identify the Boardwalk (bottom right corner) as the most profitable property, which is a logical conclusion as it is the property with the highest rent, highest base rent/cost ratio and reachable by both dice rolls and Chance. It is also observable that the railroads (properties in the middle of all four sides of the board) are relatively more profitable than other properties. The inexpensive properties on the left side of the board, despite their low cost, are revealed to not be profitable when the game is played for a long period of time.