Categorical, Time Series, and Spatial Data

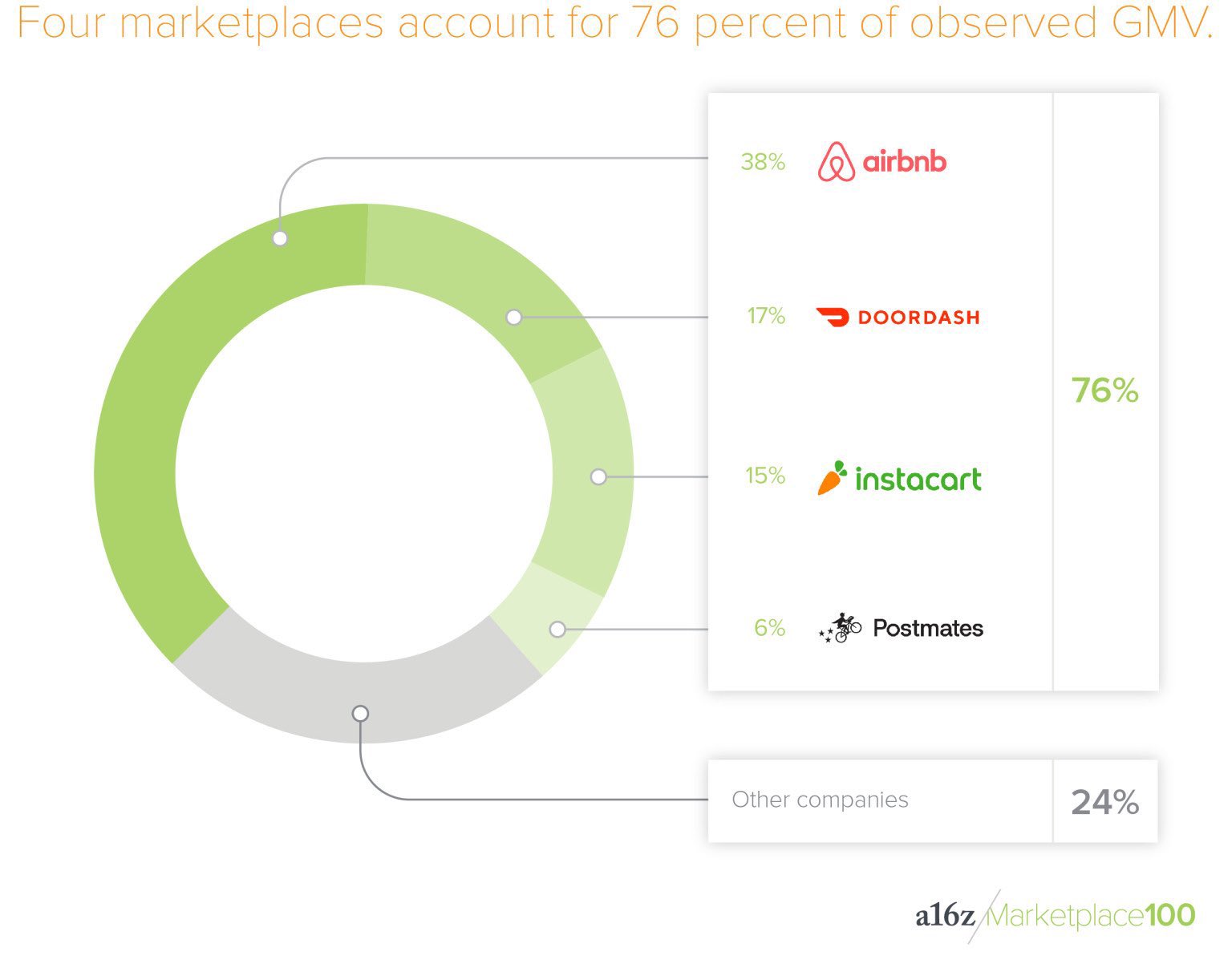
Michael Surdek

Southern New Hampshire University

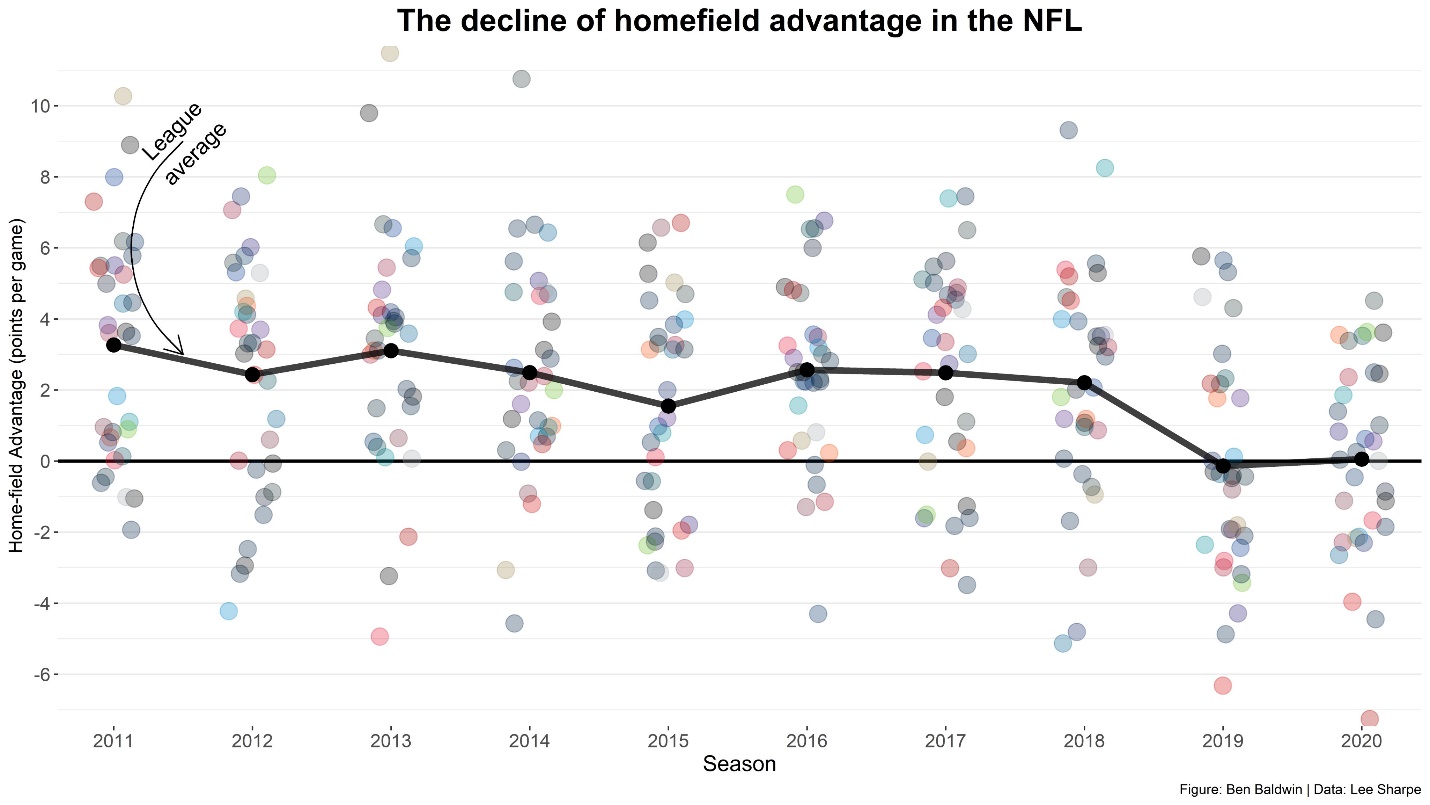
There are many effective strategies for visualizing non-numeric data including categorical, time-series, and spatial data. Data attributes that represent groups can be visualized by charts such as bar graphs, pie charts, and mosaic plots. These charts allow the audience to understand the aggregation of categorical data and easily make comparisons across groups and sub-groups. A business might need to visualize categorical data when they are analyzing sales data across multiple outlet locations. In this scenario, the outlet ID is a categorical variable and therefore each outlet could be represented by a bar that contains its sales metrics. It is easy and effective to incorporate multiple categorical variables into bar charts. In the prior example, the bars could be split into sections based on a product category, or they could be assigned a color based on the region of the United States where they are located.

Attributes that represent data over a period of time can be visualized by most chart styles where one axis is a measure of time. Lines, bars, and dots are all used to demonstrate values or the relative position of values over a period of time. A business might need to visualize time-series data if they are looking to measure their growth by quantifying their number of employees in each year over a given time period. In this scenario, bars or points would do an effective job of displaying the values over time, but it would be best to include lines connecting the data to more clearly show the changes over time. In time-series data, the changes across time are usually the most important aspect of the chart, so it is often best to include lines as either the only visual cue or on top of other visual cues such as bars and dots.

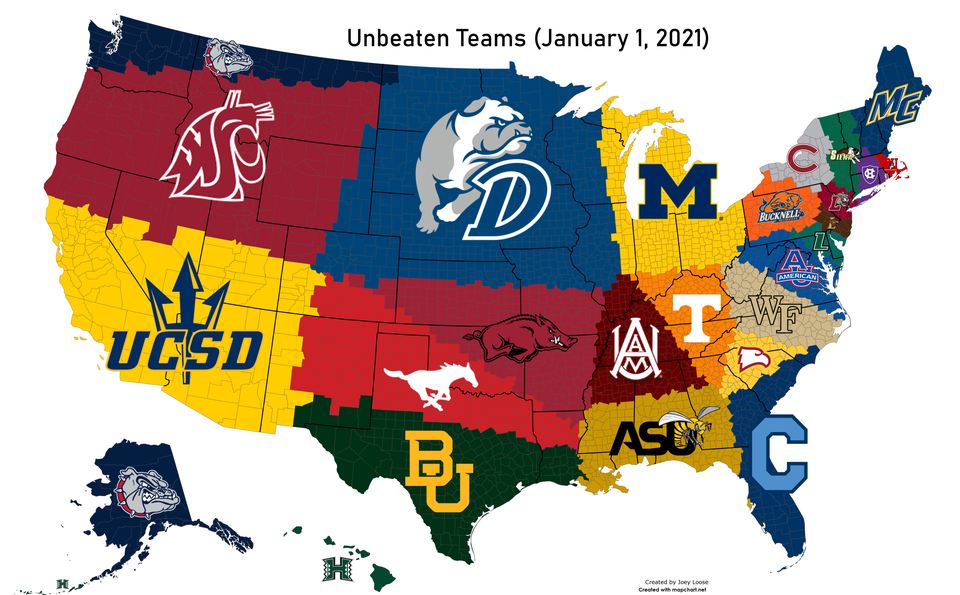
Data that contains information about where an observation took place can be visualized by many different types of mapping styles. Since humans are generally aware of the shapes of continents, countries, states and so on, it is logical to visualize spatial data by incorporating visualization techniques into familiar maps. These techniques often include alterations of the colors, sizes, and shapes of standard maps in order to represent values of other data attributes based on their location. A business might visualize spatial data on a map of the United States in order to determine where their e-commerce business is coming from and plan future distribution center development. Visualizing this type of data can provide insights that standard aggregate counts and sums cannot. In this scenario, the business could use statistics to determine locations where distribution centers would lower costs, but visualizing the spatial data would be the best way to get an idea of exactly how many distribution centers they might need to add in the future, and where exactly the ideal locations would be.



The categorical visualization example that I found shows the market share for service sharing platform start up companies. The information that the visual tries to display is that four companies account for 76% of the entire market, and all other companies only account for 24%. The category that is being displayed is the start up company such as AirBnB, Doordash, and InstaCart. Each of the companies included in this categorical variable are separate entities and together they make up 100% of the marketplace. This visual does an effective job using the visual cue of size to show the percent of the market share of each of the top four companies. The use of a separate legend allows for the company logos to be easily readable and it also provides the exact percentage of the market for each company as well as the sum of the top four vs the rest. I think the visualization would be improved if the companies and percentages were displayed within the visual itself. Although the legend makes the words more readable, it forces the audience to move their attention back and forth between the pie chart and the legend. If space were the reason that the creator made this decision, they could use a standard pie chart which would provide more space on each individual slice.



The time-series visualization example that I found shows how homefield advantage in the NFL has declined over time and all but disappeared in the last two seasons. This chart uses points that are spread across the y-axis according to the measure of a teams homefield advantage in a given season for each the last ten NFL seasons on the x-axis. Additionally, a line is connected across the chart that shows the league average and how it has changed over time. This visual is effective in showing that homefield advantage has declined over time because it can be seen by both the position and spread of the dots moving lower in more recent years and also the league average line being near 4 in 2011, near 2 from 2015-2018, and then near 0 in 2019 and 2020. The only thing I would suggest to improve this chart would be to incorporate more historical data and go back further than 2011. With more data, it would be harder to see changes between years, but it would show whether homefield advantage trending towards zero is happening for the first time or if it has happened before due to general variance or unpredictable events such as COVID-19 not allowing fans into stadiums this year for the most part.



The spatial visualization example that I found shows a map of the counties across the United States and the closest undefeated NCAA Men’s Division I Basketball program to them as of January 1st, 2021. This visualization is posted every day on the subreddit r/collegebaskeball and it is interesting to see how it changes when a previously unbeaten team loses a game. There is no real significant purpose of this visualization, other than to provide bragging rights for fans of undefeated teams. The visual is produced by using the latitudes and longitudes of US counties and unbeaten Universities to find the minimum distance between a county and any of the Universities. The map is effective in showing which programs are surrounded by other unbeaten programs, and which ones do not have other unbeaten programs anywhere near them, Washington State for example, which might take up the most space of any program. One thing that could be done to improve the visualization would be to include each program’s record. Additional numeric data can be helpful when interpreting spatial data on a map, and in this case it would show many of the “unbeaten” teams in the Northeast belong to the Patriot League Conference, which had not begun play as of January 1st, so those programs are technically 0-0.