Fundamentals of Data Visualization

A class at the
University of Applied Sciences Northwestern Switzerland

Nils Fahrni
BSc Data Science Student

June 14, 2024

Contents

Chapter 1
Introduction

In our data-driven age, the ability to visualize complex datasets not only simplifies information but also tells a story, highlights patterns, trends, and anomalies that might otherwise remain hidden. Data visualization acts as the bridge between the raw, often impenetrable world of numbers and an accessible visual representation that can be easily understood, interpreted, and acted upon from any perspective.

In this report, we delve into the fundamentals of data visualization, elucidating its importance and methodologies. By applying these principles, we present a series of visualizations centered around wildfire data from California. Through these graphics, I aim to shed light on the frequency, intensity, and geographical spread of these fires over recent years.

Join me on this journey through the five learning objectives, as we navigate the terrain of data visualization and uncover the hidden tales of California's fiery landscape — a landscape where not only trees and vegetation are at risk, but also a realm of danger for its myriad forest animals.

The data utilized in this report has been sourced from California's Department of Forestry and Fire Protection, available at www.fire.ca.gov. This comprehensive dataset covers an eight-year span, from 2013 to 2020.

The project repository of this report can be found at github.com/okaynils/fhnw-ds-gdv.

Chapter 2
Visualization basics and chart types

We aren't innately equipped with the knowledge to interpret any type of chart which displays data. As creators of data visualizations, we need to ask ourselves the question of "Who is the visualization at hand for?". It's crucial for us to comprehend our audience's perspective and recognize when an alternative graph might captivate them more effectively or be easier on their eyes, aiding in enhancing our audience's graphical literacy.

Scott Klein, deputy managing editor at *ProPublica* once wrote,

> "There is no such thing as an innately intuitive graphic. None of us are born literate inreading visualizations."

This realization brings light to the fact that we cannot create visualizations that need absolutely no thinking. There will always be a subjective interpretation — but, as a data scientist, one can help to make the process of interpretation easier and more streamlined for those who have to interpret the insights.

Visualizing frequencies

A great way to visualize frequencies of nominal data is to use **bar charts**. In this example we visualize the wildfire season's frequency throughout the average year.

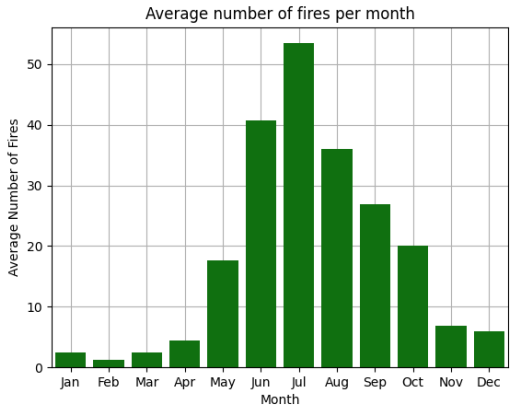

Figure 2.1: Bar Chart

When to use bar charts

Bar charts are among the simpler types of charts. They allow us to get an overview of how frequent a category is in a quick manner. It not only shows us a single category, but displays all the categories at once, allowing us to compare the magnitudes.

When to not use bar charts

Bar charts can have a few flaws. It is important to notice that bar charts work best if there aren't too many categories present. Such a chart can also get over-cluttered fast if too much information is added, which may not be the result one wants.

Comparing densities

A violin plot is a method of plotting numeric data and can be understood as a combination of a box plot and a kernel density plot. It provides a visualization of the distribution of the data, its probability density, and its cumulative distribution, with the added ability to visualize the data's symmetrical nature and the presence of multiple modes.

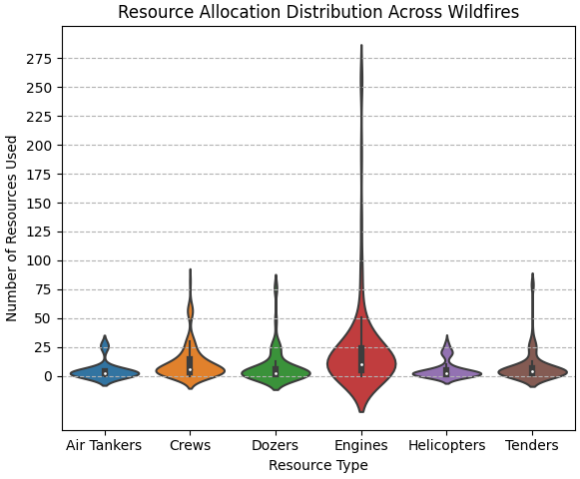

Figure 2.2: Violin Plot

When to use Violin Plots

Violin plots excel at visualizing data distributions across categories. For the wildfire dataset, they aptly showcase how resource deployments vary

across different fire conditions, revealing both average usage and distribution nuances.

When to avoid using Violin Plots

Violin plots may not be the best choice when the dataset has limited data points or when the primary interest is in summary statistics rather than detailed distributions. For instance, in the wildfire context, if we're solely focused on the average number of helicopters used per year, a simple bar chart would be more straightforward and interpretable. Generally, Violin Plots are said to be on the harder side when it comes to interpretability.

Geospatial Visualization

In this example we have location based data of wildfires at our hand. If we want to know which Californian counties were most heavily affected, this view offers a great opportunity to visualize that. This following visualization gives us a location based overview on how many acres were burned in each Californian region.

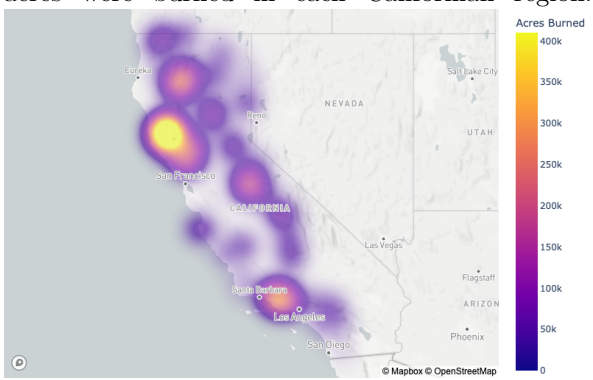

Figure 2.3: Heatmap

The problem with geospatial visualizations

Something we notice when looking at the heatmap is that we only get a rough visual estimate on how large the fires actually were. The cells on the heatmap don't represent the actual area that was burned, only their color does. The cells start to get bigger when other fires are in close proximity. This can lead to misinterpretation.

When to use geospatial heatmaps

These visualizations are often used to show an event's severity and location. Such heatmaps generally work well if there should be an immediate visual impact on the viewer as humans can quickly identify areas of severity thanks to *heat bubbles*.

When to avoid using geospatial heatmaps

As already touched on in a previous section, these visualizations often don't work well when wanting to show the exact extent of an event. The *heat bubbles* show a gradient of colors — This makes it hard for users to pinpoint the actual value belonging to the color on the legend.
Heatmaps often give the perception of continuous data, but in this case, the data (acres burned) is discrete to each county. The smooth transitions between colors could suggest a continuity that doesn't exist in the data. So users might lean towards thinking that the bubbles show the area of burning forest.

Chapter 3
Visual Perception

Making extremes stand out

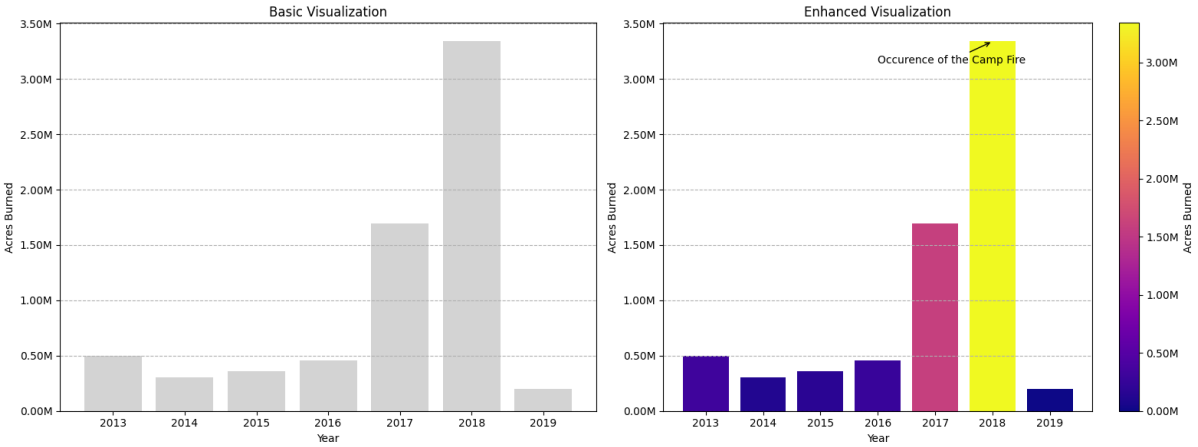

Figure 3.1: Color enhancement comparison

On the left, we have the **Basic Visualization** without colors and annotations, using a uniform light gray color for all bars.

On the right, we have the **Enhanced Visualization** with color gradients representing the intensity of acres burned, the year with the maximum acres burned highlighted in red, and an annotation providing context about the maximum value.

Of course one could say the first "Basic Visualization" is more than enough but the latter version even allows to capture the year of the most damage in a more dramatic way without having to examine axes and context closer.

3D and its perceptual problems

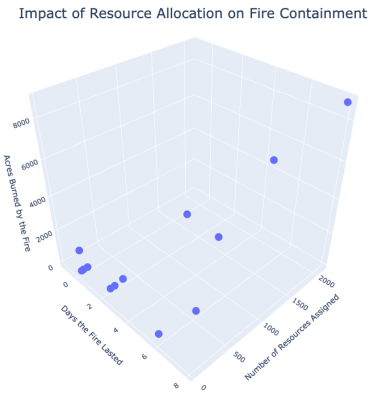

Figure 3.2: 3D Plot

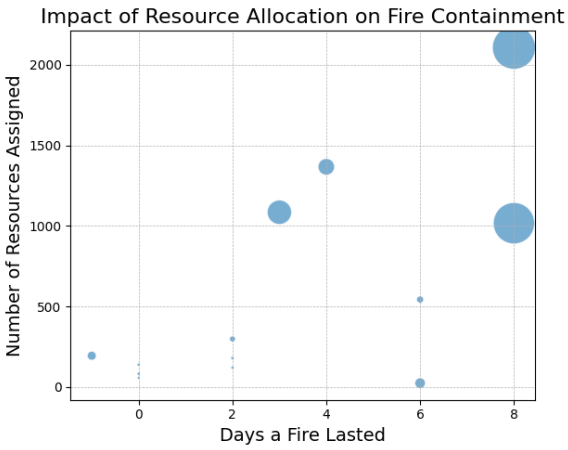

Figure 3.3: Simplified Plot

2D plots are often preferred for clarity and simplicity. They can be easier to read, especially when printed or viewed on a flat medium like paper or a non-interactive display. 3D plots, on the other hand, can provide additional perspective and are excellent for displaying complex datasets with three distinct variables. However, without interactivity, 3D plots can become confusing and may not effectively communicate the data due to issues with perspective, such as occlusion (where parts of the graph obscure other parts) or distortion (where the angle of view can make certain dimensions appear different than they are). Here, the 2D plot shows the exact same data as the 3D plot but we made use of another property, the **scatter size**.

Counteracting Overlapping Data

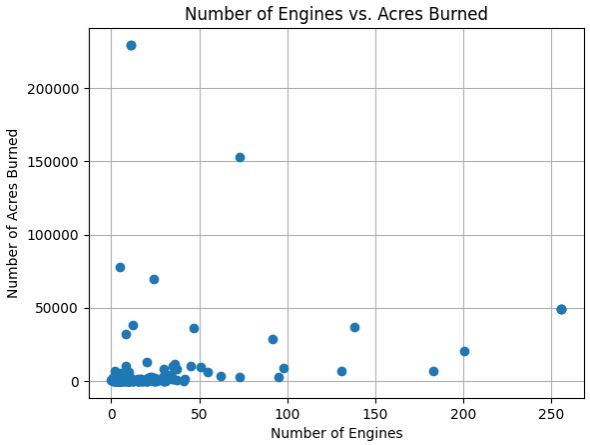

Figure 3.4: Default Scatter Plot

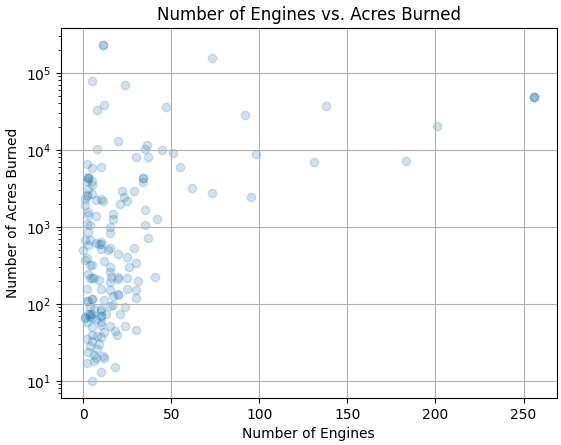

Figure 3.5: Enhanced Scatter Plot

The data points are clustered at the lower end of the engines axis, indicating that most observations have a relatively low number of engines involved. Overlapping data points, particularly where the concentration of points is high, can obscure individual data values and may indicate a need for more sophisticated plotting techniques. Let's see how we can modify the **scatter plot** to see what actually lies behind the dense cluster of points.

Applying a lower alpha value to the scatter points, as shown in the right plot, helps to mitigate the issue of overlapping data points by making the points semi-transparent, which allows for the visualization of density in regions with high data point concentration. Furthermore, transforming the y-axis to a logarithmic scale spreads out the points along the vertical axis, providing a clearer view of the data across a wider range of values. However, this approach can make it more difficult to interpret the actual values, especially for those unfamiliar with logarithmic scales, as the scale compresses the data range and changes the perception of the distances between points. The reduced emphasis on individual, more transparent points can make it challenging to notice outliers or standalone points which could be significant. This trade-off in clarity for density versus emphasis on individual data points requires careful consideration when visualizing data to ensure accurate interpretation.

Chapter 4
Design Principles vs. Data

In this chapter, we embark on a journey through the various stages of data visualization, focusing on geospatial data and its visualization. The goal of this demonstration is to create a **choropleth map** of California's counties, showing the most frequently affected counties when it comes to wildfires.

Data Preparation

The goal of the first step of creating such a visualization is to decide what kind of data we need to display our desired graphic.

- Calculate the fire frequency for each county: The choropleth map is going to color in the county's areas according to how many fires the data set has recorded.
- Next to the counties and their frequencies of wildfires we will be required to aggregate their locations on the map. The data set we have at hand provides us with latitude and longitude.
- Perhaps not every county is represented in the data set – We therefore need to make sure to add these counties with zero incidents so the desired geomapping framework (Mapbox) can handle our needs.

Select and Tailor Important Features

After preparing the data on a general level the data needs to be cropped to its needs.

- Implement a disambiguation process by cross-referencing county names with state identifiers. In the case of counties like "Lake County", which may exist in multiple states, it is crucial to include a state-level filter in the data processing pipeline. This ensures that only the "Lake County" located within the boundaries of California is included in the dataset used for the choropleth map.
- Cross-check the county names and their associated data against a reliable source, such as a government geographical database or API. This step can help to confirm that the data being visualized corresponds correctly to the intended locations within California.

Following Design Principles

Choosing Appropriate Colors

Effective choropleth maps for visualizing Californian wildfires should highlight accurate and meaningful spatial patterns, ensuring clarity and readability both in grayscale print and colored digital displays. Moreover, these maps must be crafted with colorblind-friendly design principles, ensuring accurate interpretation by a diverse audience, including those with color vision deficiencies. This approach guarantees that the maps are not only informative but also inclusive, providing a comprehensive understanding of wildfire distribution and impact across California [1].

In this case the sequential color map "plasma" was used which is one of the out-of-the-box available colormaps from matplotlib. These default colormaps were made with all the common color design principles in mind, including colorblindness. The color map is also perceptually uniform, meaning that the colors are evenly distributed across the color spectrum. This is important because it ensures that the colors are not biased towards a certain color, which could lead to misinterpretation of the data [2]. The chosen color map generally follows a linear sequential pattern which is what we want in our case as visual hierarchy is important when wanting to show frequency magnitudes of fires.

The Design Principles of Legends

These following few points are important to keep in mind when designing a legend for a choropleth map:

- The placement of the legend is crucial for the viewer to easily associate the labels with the corresponding data. Placing the legend below or parallel to the visualization ensures a clear connection between the two [3].
- The decision on whether to include a legend title is based on the need to provide additional context to the visualization. If the title adds value and clarity to the chart, it should be included; otherwise, it can be omitted [3].
- Directly labeling data representations instead of using a legend can enhance the understanding of the chart, as it reduces the cognitive load on the viewer. This principle aims to make the chart more intuitive and easier to interpret [4]. In the case of the choropleth map it is not possible to directly label the data in an aesthetically pleasing way as some counties are too small to fit the label in.

Showing Correct Scales

The Choropleth map we try to visualize in this chapter should show which counties or therefore regions are most critically affected by wildfires. Depending on this task we need to think about how we want to scale the data.

A common conception would be to assume that larger counties naturally have more wildfires, though that may not be true – A county may be large but if the overall forest coverage is low, the rate of fires would most likely also be low. To counteract these underlying causalities we could scale the data [5]: In this case it would mean to divide the number of fires by the area of forest and thus display the rate of fires per acre through our color palette.

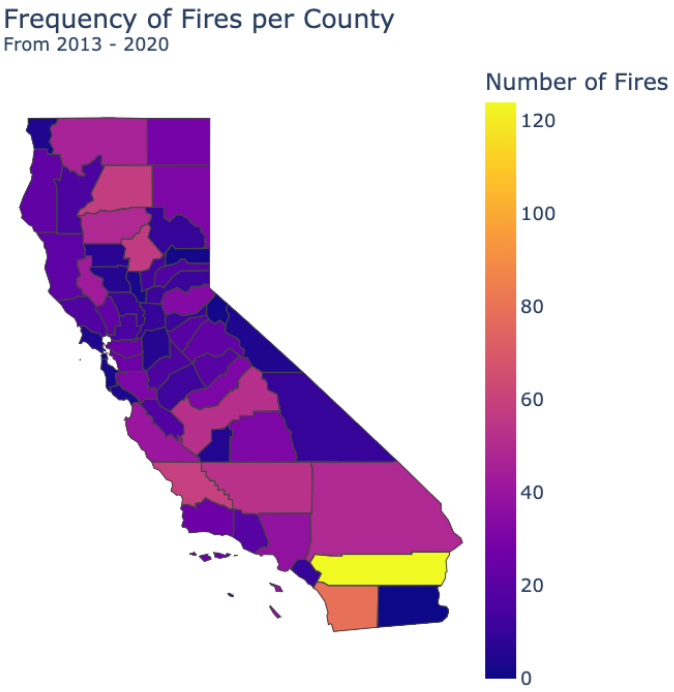

Figure 4.1: Choropleth

Chapter 5
Grammar of Graphics Tools

The Grammar of Graphics by Leland Wilkinson is a framework that provides a unique foundation for producing almost every quantitative graphic found in scientific journals, newspapers, statistical packages, and data visualization systems.

The Layers: An Overview

The Grammar of Graphics follows a layered approach to describe and construct visualizations or graphics. It provides a common language for thinking about the ways that design choices are made in visualization, describing everything from the data used to the visual channels displayed on the marks, and how data is converted into those channels.

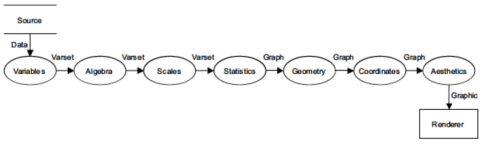

Figure 5.1: Grammar of Graphics Flow [6]

Data

In the first step, as described above, we aggregate and transform the data into a format we can later use to visualize and perform further visual modifications on. If we visualize the data at this step, we can see that the plot doesn't show much meaningful information yet.

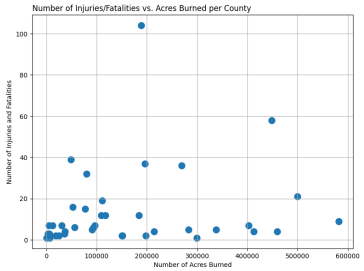

Figure 5.2: Grammar of Graphics: Data

Aesthetics

In the Grammar of Graphics, the Aestehetics Layer refers to the mapping of one or more variables to one or more visual elements on the graph. This includes mapping variables to the x-axis, y-axis, and using color to differentiate different attributes. Aesthetics are essential in creating meaningful and effective visualizations [7]. In this example we map the `AcresBurned` variable to the x-axis and the sum of `Injuries` and `Fatalities` variables to the y-axis. Additionally, the Admin Units (fire departments) get added as a third dimension by using color.

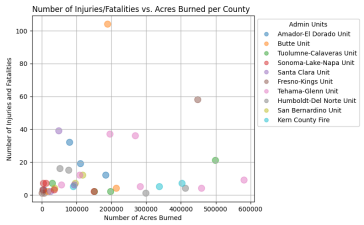

Figure 5.3: Grammar of Graphics: Aesthetics

Scale

In the Grammar of Graphics, the Scale step can include the scaling of the data, but also the scaling of the visual elements, such as the axes or scatter sizes [8]. In this example we added another dimension to the plot by utilizing the scale of scatters (Fire Dept. Size) to visualize the number personnel that was involved. Additionally, since most of the scatters were overlapping in the previous linear x-scale, we switched to a logarithmic scale for the x-axis.

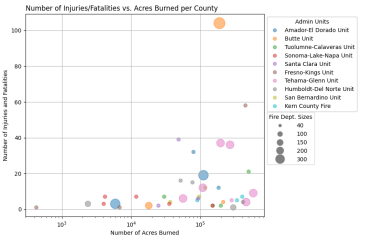

Figure 5.4: Grammar of Graphics: Scale

Geometry

We can use the Geometry Layer of the Grammar of Graphics to access further dimensions without having to add more axes in the sense of a 3D plot [9]. In this example we use the shape-property of the scatterplot and its scatters to visualize the fuel types that accelerated a wildfire. As each scatter represents a county the shape of the scatter describes the county's major fuel type; The type of fuel that is responsible the most in a given county.

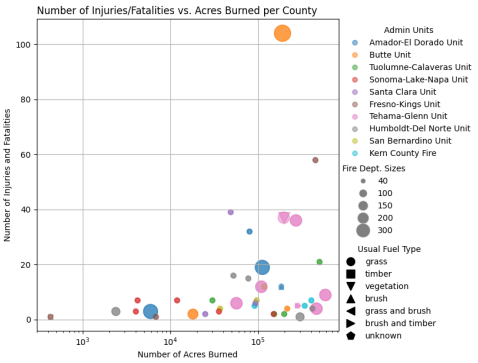

Figure 5.5: Grammar of Graphics: Geometry

Statistics

Showing statistical features in a visualization helps the viewer to form a better contextual understanding as for example the mean provides a good reference point for an "Expected Value" or the general center of data with the intersection of the `vline` and `hline` [10].
Other types of visualizations could include a confidence interval or a regression line as another statistical reference point.

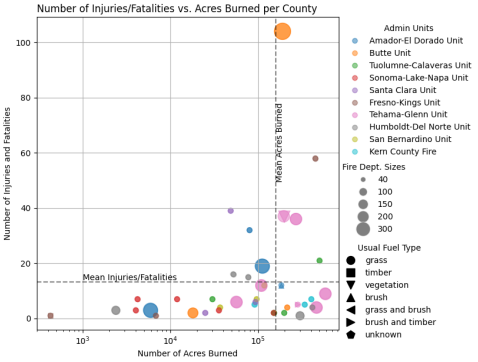

Figure 5.6: Grammar of Graphics: Statistics

Facets

Facets in the Grammar of Graphics are a way to split data into subplots based on another factor in the data. They allow for the creation of multiple small multiples, each showing a different subset of the data. In this case, the plot we are working on was split into five subplots, one for each category of fire department size. The respective Admin Unit (fire dept.) of each county has a size which represents how many personnel were involved in the fire.

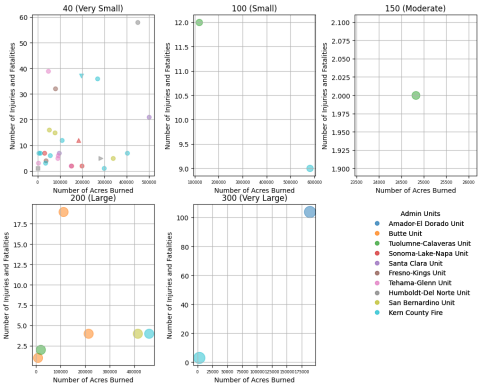

Figure 5.7: Grammar of Graphics: Facets

Coordinates

When looking at coordinates in the context of the Grammar of Graphics, we are referring to the coordinate system of the plot [11]. This includes the type of coordinate system, such as Cartesian or Polar. It also includes the range of the coordinate system, such as the range of the x-axis and y-axis.

In this case, we have implemented a Cartesian coordinate system with a logarithmic x-axis. Choosing what kind of coordinate system to use mainly depends on the type of the underlying data. The most common coordinate system is the Cartesian coordinate system as it is very intuitive for human perception. Polar coordinates on the other hand are hard to comprehend as humans usually perform rather bad when it comes to interpreting angles, which is the main component of polar coordinates [12, Chapter 7].

Chapter 6
Evaluation

Evaluating visualizations is of paramount importance because it impacts how individuals, organizations, and societies interpret and make decisions based on data. Poorly designed visualizations can lead to misunderstandings, incorrect conclusions, and misguided decisions, whereas effective visualizations can facilitate better understanding, insights, and actions. In fields where data and decisions are critical—such as science, business, and public policy – the ability to accurately and effectively visualize information is a powerful tool for communication and understanding.

Evaluating Visualizations with Usability Testing

Usability testing is a popular User Experience research methodology that focuses on collecting insights, findings, and anecdotes about how people view a visualization. In a usability-testing session or survey, a researcher asks participants to perform tasks and interpret specific visualizations while observing the participant's behavior and listening to their feedback. Qualitative usability testing is best for discovering problems in the user experience and is more common than quantitative usability testing, which focuses on collecting metrics to describe the user experience. It's important to perform such a survey with more than just one sample person as the goal mainly is to find overlapping and individual observations and patterns in the feedback which both can be used to either validate or improve a visualization [13].

In this report, these following two visualizations will be evaluated with a usability test (survey):

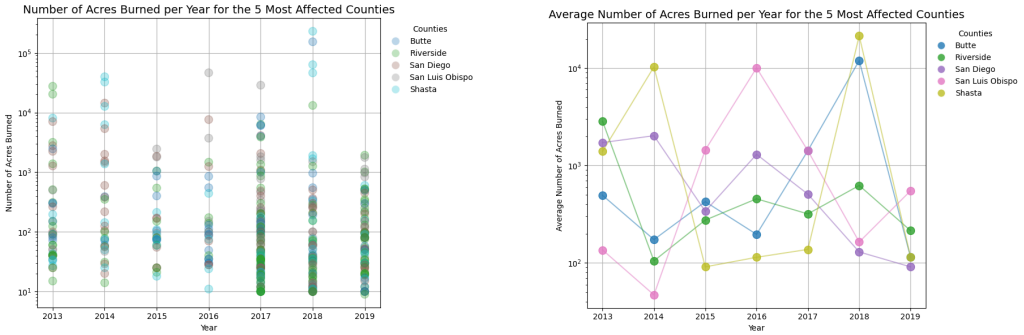

Figure 6.1: Visualizations used in the Usability Test

Defining the Usability Test's Environment

For the usability test of the visualizations, a cohort of five individuals from diverse backgrounds, ages and professions were selected to ensure a wide range of perspectives and experiences.

The following questions were asked for both visualizations:

#	**Question**
1	Which county seems to have been affected the most severe by fires?
2	Which year was the worst when it comes to wildfires?
3	Are there any elements or aspects of the visualization that confuse you?
4	Can you see a relationship between the year and the acres burned?
5	Can you see any pattern for a single county, if so what can you notice?

Table 6.1: Usability Test Questions for Visualizations

The questions were asked during a video call with the participants. The visualizations were given to the participants in random order but one at a time.

Analyzing the Results

Question 1

The first visualization consistently indicates that Butte County has been most severely affected by wildfires, particularly in the year 2018. This observation is clear for 4 of the 5 subjects, showing that the visual distinction of Butte County's data is effectively communicated in the design. However subjects that elaborated on the second visualization correctly noted that it is not clear which county was most affected as the data points got averaged.

Question 2

All subjects identified 2018 as the year with the most significant impact from wildfires in both visualizations. This validates the visualizations' effectiveness in conveying temporal trends and highlighting outlier events, which is crucial for understanding the temporal scope of the data.

Question 3

4 of 5 subjects noted potential confusion arising from the log scale on the y-axis and the overlapping data points in the first visualization. The second visualization's line graph was also noted to potentially imply continuity where there is none. Thus, these points resemble areas for design refinement to enhance the clarity and a more intuitive understanding. All five subjects struggled with the question when looking at how much time they needed – An observation that seems to have caused this was that most of the subjects were trying tell what number was lying beneath the y-scale.

Question 4

The relationship between the year and the number of acres burned is not consistently apparent across subjects, with some recognizing an increasing trend, while others see variability. 3 of the 5 subjects tried to name a trend by mainly looking at the connecting lines in the second visualization. The goal of the lines between the scatters was to make it easier to follow each county over time – This seems to have caused confusion and would need to be improved in a future iteration.

Question 5

Patterns within single counties were observed differently by all subjects, with some noticing fluctuating trends and others identifying specific years of extreme wildfire activity. The variance in interpretation leads to the fact that while individual events are highlighted effectively, the overall pattern communication could be improved to ensure consistent comprehension among all viewers.

Summary

The goal with these visualizations was to convey the severity of wildfires in California for the biggest counties. All of the subjects were focusing on the actual number of acres which were burned and not the overall trend and severity. The biggest reason for that contextual misunderstanding was the log scale on the y-axis. The extremes were pointed out correctly by all subjects and a trend was not falsely identified by anyone.

Bibliography

[1] Jack Dougherty and Ilya Ilyankou. Design Choropleth Colors & Intervals. In *Hands-On Data Visualization*. Picturedigits Ltd., 2021.

[2] L.D. Bergman, B.E. Rogowitz, and L.A. Treinish. A rule-based tool for assisting colormap selection. In *Proceedings Visualization '95*, pages 118–125,, Atlanta, GA, USA, 1995. IEEE Comput. Soc. Press.

[3] Legends | Data Visualization Standards. https://xdgov.github.io/components/legends.

[4] Carbon Design System. https://www.carbondesignsystem.com.

[5] Jack Dougherty and Ilya Ilyankou. Normalize Choropleth Map Data. In *Hands-On Data Visualization*. Picturedigits Ltd., 2021.

[6] How To Make a Pie. In Leland Wilkinson, editor, *The Grammar of Graphics*, Statistics and Computing, pages 23–40. Springer, New York, NY, 2005.

[7] Aesthetics. In Leland Wilkinson, editor, *The Grammar of Graphics*, Statistics and Computing, pages 255–318. Springer, New York, NY, 2005.

[8] Scales. In Leland Wilkinson, editor, *The Grammar of Graphics*, Statistics and Computing, pages 85–109. Springer, New York, NY, 2005.

[9] Geometry. In Leland Wilkinson, editor, *The Grammar of Graphics*, Statistics and Computing, pages 155–178. Springer, New York, NY, 2005.

[10] Statistics. In Leland Wilkinson, editor, *The Grammar of Graphics*, Statistics and Computing, pages 111–154. Springer, New York, NY, 2005.

[11] Coordinates. In Leland Wilkinson, editor, *The Grammar of Graphics*, Statistics and Computing, pages 179–254. Springer, New York, NY, 2005.

[12] Fletcher. Dunn and Ian. Parberry. 3D math primer for graphics and game development. In *3D Math Primer for Graphics and Game Development*. CRC Press, Boca Raton, Fla, 2nd ed. edition, 2011.

[13] NNgroup. Usability Testing w. 5 Users: Design Process (video 1 of 3), October 2018.